COURSE GUIDE

PHS 421
RESEARCH METHODS IN PUBLIC HEALTH

Course Team Dr. Joseph O. Ajobiewe & Dr. Abioye J.O.K. (Course Writers)
Dr. Gloria Anetor (Course Coordinator) - NOUN

Prof. Grace C. Okoli (Programme Leader) - NOUN

NATIONAL OPEN UNIVERSITY OF NIGERIA

National Open University of Nigeria
Headquarters
University Village
Plot 91, Cadastral Zone
Nnamdi Azikiwe Expressway
Jabi, Abuja

Lagos Office
14/16 Ahmadu Bello Way
Victoria Island, Lagos

e-mail: centralinfo@nou.edu.ng
URL: www.nou.edu.ng

Printed 2019

ISBN: 978-978-970-099-8

CONTENTS **PAGE**

INTRODUCTION

The course, PHS 421, takes you into the basic characteristics of research. Here, we shall take an in-depth consideration of the meaning of research, objectives of research, and various types of research; such as Descriptive, Analytical, Applied, Fundamental, Quantitative, Qualitative, Conceptual, Empirical, Cross-sectional, and Longitudinal Researches. We shall focus on knowledge and skills in planning and conducting Public Health research including behavioural, epidemiological and health systems studies. You will also be exposed to the nature and uses of research in health, research processes and designs, methods of data collection, data analysis and presentation, research proposal and report writing.

WHAT YOU WILL LEARN IN THIS COURSE

We shall learn the following briefly in the process of our working through Research methods in public health; we shall look at, Meaning of Research, Objectives of research, Types of Research, Research approaches, Significance of research, Research Methods, Research Process. Defining research problem, viz; - what is a research problem? Selecting the problem; Necessity of defining the problem; Technique involved in defining a problem;

Research Design, meaning of research Design; Need for research Design; Features of a good Design; Important Concepts Relating to research Design, viz; - Dependent and independent variables, Extraneous variables, Controls, Confounded Relationships, Research Hypothesis, Experimental and Non –experimental Hypothesis-Testing Research, Experimental and Control Groups, Treatments, Experiment, Experimental units, Different Research Design; Basic Principles of Experimental Designs.

Data Collection; Experiments and Surveys, Collection of primary data, Collection of secondary data Selection of appropriate Method for Data Collection.

Data Preparation; Data preparation process viz; - Questionnaire Checking, Editing, Coding, Classification, Tabulation, Graphical Representation, Data cleaning, Data Adjusting, Missing Values and Outliers, Various Types of Analysis as related to Public Health, such as, Multiple Regression Analysis, Multiple discriminate Analysis, Multiple Analysis of Variance(ANOVA), Canonical Analysis and Inferential Analysis.

Report Writing; Significance of report Writing, Different Steps in Writing Report, Layout of the Research Report, Types of Reports, Oral Presentation, Mechanics of Writing a Research Report, Precautions for Writing Research Reports.

After this, we shall practically illustrate all the aspects stated above in the forms of bio- and vital statistics related to public and environmental health.

COURSE AIM

The aim of this course is to provide a good understanding of Public Health research methods such as to gain familiarity with a phenomenon or insights into it, as is done in public health exploratory or formulative research.

Other aims include:

- to portray accurately the characteristics of a particular individual, situation or a group playing significant role/s in a public health descriptive research instance
- to determine the frequency with which something occurs or with which it is associated with something else in public health
- to test a hypothesis of a causal relationship between variables of interest in a particular public health situation.

COURSE OBJECTIVES

By the end of this course, you will be able to:

- explain the meaning of research
- discuss the objectives of research
- list types of research
- discuss research approaches
- explain the significance of research
- explain research Methods
- define research problem, viz; -what is a research problem? Selecting the problem; necessity of defining the problem; technique involved in defining a problem
- explain research design in detail
- discuss the different research design
- enumerate the steps involve in data collection
- explain data preparation
- discuss the different steps involved in writing research report.

WORKING THROUGH THIS COURSE

This course has been carefully put together bearing in mind that you might be new to the course. However, efforts have been made to ensure that adequate explanation and illustrations were made to enhance better understanding of the course. You are therefore advised to spend quality time to study this course and ensure that you attend tutorial sessions where you can ask questions and compare your knowledge with that of your course mates.

COURSE MATERIALS

The major components of the course are:

- Course Guide
- Study Units
- Textbooks
- Assignment File
- Presentation Schedule

STUDY UNITS

This course comprises five modules, broken down into 12units. Theyare listed as follows:

Module 1

Unit 1	Meaning of Research
Unit 2	Other Research Features and Characteristics
Unit 3	Sampling

Module 2

Unit 1	Data Collection
Unit 2	Population Census
Unit 3	Life Table

Module 3

Unit 1	Data Preparation
Unit 2	Public Health Statistics

Module 4

Unit1 Further Features of Morbidity and Mortality Indices
Unit 2 Summary of Biostatistics

Module 5

Unit 1 Writing Report
Unit 2 Proposal Writing

In module one, unit one shall introduce us to the Meaning of research, unit two to other features of research and its characteristics.

In Module two, we shall consider Data Collection in unit one; in unit two, we shall briefly look at the population census as typical cross sectional research.

In module three, unit one introduces you to Data presentation, unit two takes us to public health statistics.

In module four, unit one, we shall consider further features of morbidity and mortality indices; in unit two we shall look at general biostatistics in public health.

In Module five, unit one takes you through writing report; unit two introduces you to proposal writing.

TEXTBOOKS AND REFERENCES

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d.). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

ASSIGNMENT FILE

In this file, you will find all the details of the work you must submit to your tutor for marking. The marks you obtain from the assignments will count towards the final mark you obtain for this course. Further information on assignments will be found in the Assignment File itself and later in this Course Guide in the section on assessment.

ASSESSMENT

There are two aspects to the assessment of the course. First are the tutor-marked assignments; second, is the written examination. In tackling the assignments, you are expected to apply information and knowledge acquired during this course. The assignments must be submitted to your tutor for formal assessment in accordance with the deadlines stated in the Assignment File. The work you submit to your tutor for assessment will count for 30 percent of your total course mark.

At the end of the course, you will need to sit for a final examination. This will also count for 70 percent of your total course-mark.

TUTOR-MARKEDASSIGNMENT

You should be able to complete your assignments from the information and materials in your set textbooks and study units. However, you are advised to use other references to broaden your view point and provide a deeper understanding of the subject.

When you have completed each assignment, send it to your tutor. Make sure that each assignment reaches your tutor on or before the deadline given. If however you cannot complete your work on time, contact your tutor before the assignment is due, to discuss the possibility of an extension.

FINAL EXAMINATION AND GRADING

The examination will consist of questions which reflect the type of self-testing, practice exercises and tutor-marked problems you have come across. All areas of the course will be assessed.

You are advised to revise the entire course after studying the last unit before you sit for the examination. You will find it useful to review your tutor-marked assignment and the comments of your tutor on them before the final examination.

SUMMARY

The Course Guide gives you an overview of what to expect and what to do in the course of this study. The course teaches you about research methodology

We wish you success in this course and hope that you will find it both interesting and rewarding.

MAIN COURSE

CONTENTS PAGE

MODULE 1

UNIT 1 MEANING OF RESEARCH

CONTENTS

1.0 INTRODUCTION

Research in common parlance refers to a search for knowledge. One can also define research as a scientific and systematic search for pertinent information on a specific topic. In fact, research is an art of scientific investigation. Dictionary definition of research is a careful investigation or inquiry especially through search for new facts in any branch of knowledge. Some people consider research as a movement from the known to the unknown. It is actually a voyage of discovery. We all possess the vital instinct of inquisitiveness. When the unknown confronts us, more and more our inquisitiveness makes us probe and attain understanding of the unknown. This inquisitiveness is the mother of all knowledge and the method, which one employs for obtaining the knowledge of whatever the unknown, can be termed as research. Research is an academic activity and as such the term should be used in a technical sense. According to Clifford woody, research comprises redefining problems, formulating hypothesis or suggested solution; and at last carefully testing the conclusion to determine whether they fit the formulating hypothesis. D. Slesinger and M. Stephenson in the encyclopaedia of social sciences define research as" the manipulation of things, concepts, or symbols for the purpose of generalising to extend, correct or verify knowledge, whether that knowledge aids in construction of theory or in the practice of an act." Research is, thus, an original contribution to the existing stock of knowledge through objective and systematic method of finding solution to a problem is research. As such the term "research" refers to systematic method consisting of enunciating the problem, formulating a hypothesis, collecting the facts or data, analysing the facts and reaching certain conclusions either in the form of solution (s) towards the concerned problem or in certain generalisations for some theoretical formulation.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- state the meaning of the word "research**"**
- explain the meaning of descriptive and analytic research
- tell us about applied vs. fundamental research
- explain quantitative vs. qualitative research
- describe conceptual vs. empirical research
- describe longitudinal research
- explain concisely the term cohort study/research.

3.0 MAIN CONTENT

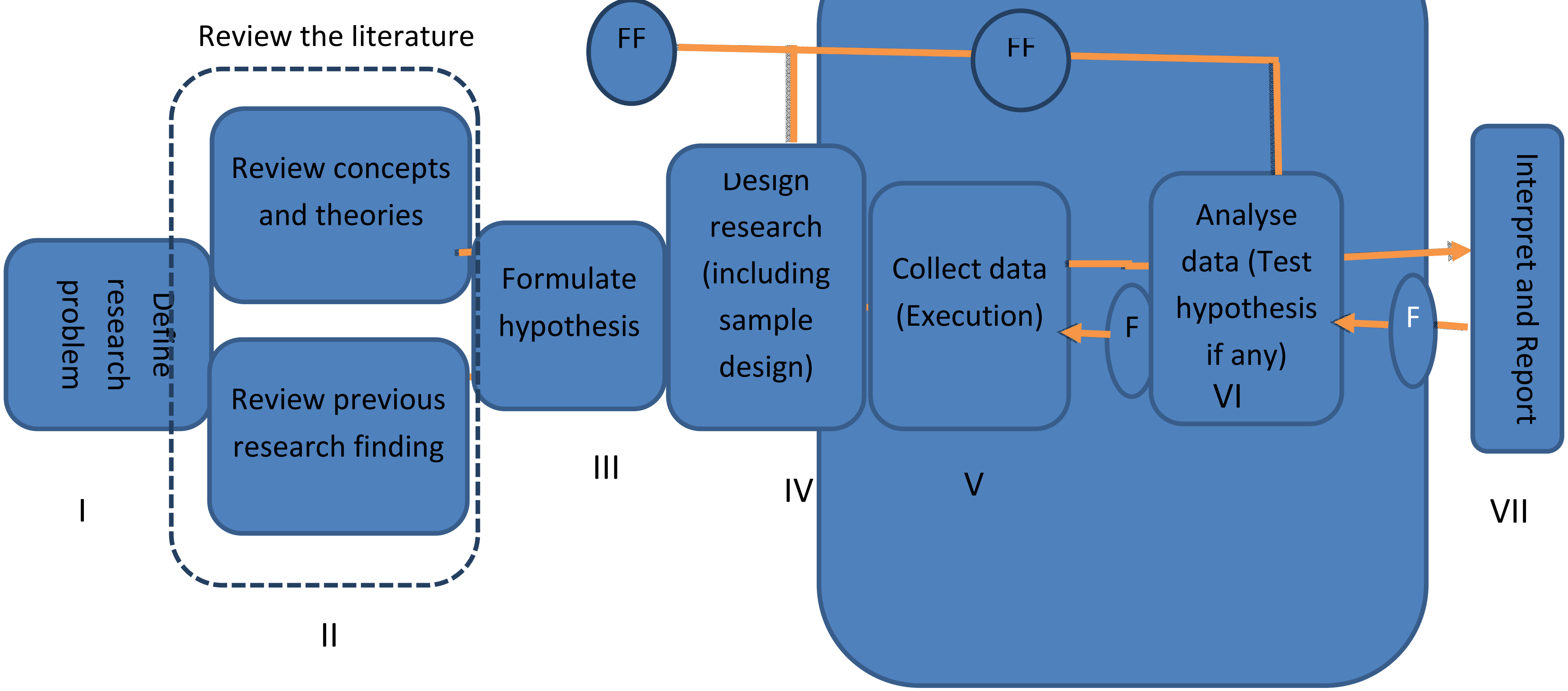
3.1 Research Process

The chart indicates that the research process consists of a number of a number of closely related activities, as shown through I to VII. But such activities overlap continuously rather than following a strictly prescribed sequence. At times, the first step determines the nature of the last step to be undertaken. If subsequent procedures have not been taken into account in the early stages, serious difficulties may arise which may even prevent the completion of the study. One should remember that neither various steps involved in a research process are mutually exclusive; nor they are separate and distinct. They do not necessarily follow each other in any specific order and the researcher has to be constantly anticipating at each step in the research process the requirements of the subsequent steps. However, the following order concerning various steps provides a useful procedural guideline regarding the research process:

i. Formulating the research problem;
ii. Extensive literature survey;
iii. Developing the hypothesis;
iv. Preparing the research design;
v. Determining sample design;
vi. Collecting the data;
vii. Execution of the project;
viii. Analysis of data;
ix. Hypothesis testing;
x. Generalisations and interpretation;
xi. Preparation of the report or presentation of the results, i.e., formal write-up of conclusions reached.

A brief description of the above stated steps will be helpful as we shall see in the later parts of this study unit.

RESEARCH PROCESS IN FLOW CHART



Where = feedback (Helps in controlling the sub-system to which it is transmitted)

F

= feed forward (Serves the vital function of providing criteria for evacuation)

3.2 Descriptive vs. Analytic Research

Descriptive research includes surveys and fact-finding enquires of different kinds. The major purpose of descriptive research in description of the state of affairs as it exists at present, in social science and business research we quite often use the term (ex post facto) research for descriptive research studies. The main characteristic of this method is that the researcher has no control over the variables; he can only report what has happened or what is happening. Most ex post factors, research on projects that are used for descriptive studies in which the researcher seem to measure such items as, for example, frequency of shopping, preferences of people or similar data. Ex post facto studies also attempt researchers to discover causes even when they cannot control the variables. The methods of research utilised in descriptive research are survey methods. In analytical research on the other hand, the researcher has to use facts or information already available, and analyse these to make a critical evaluation of the material.

3.3 Applied vs Fundamental Research

Applied research aims at finding a solution for an immediate problem facing a society or an industrial/business organisation, whereas fundamental research is mainly concerned with generalisation and with the formulation of a theory. Gathering knowledge's sake is termed fundamental research. Research concerning some natural phenomenon or relating to pure mathematics are examples of fundamental research. Similarly, research studies, concerning human behaviour carried on with a view to make generalisations about human behaviour, are also examples of fundamental research. However, research aimed at certain conclusions facing a concrete social, economic or political trends that may affect a particular institution, marketing research, evaluation research are examples of applied research. Thus, the central aim of applied research is to discover a solution for some pressing practical problems, whereas basic research is directed towards finding information that has a board base of applications and thus, adds to the already existing organised body of scientific knowledge.

3.4 Quantitative vs Qualitative Research

Quantitative research is based on the qualitative measurement of some characteristics. It is applicable to phenomena that can be expressed in terms of quantities. Qualitative research, on the other hand, is concerned with qualitative phenomena i.e., phenomena relating to or involving qualitative or kind. For instance, when we are interested in investigating the reasons for human behaviour (i.e. why people think or do certain things), we quite often talk of (Motivation Research) an important type qualitative research. This type of research is aim at discovering the underlying motives and desires, using in depth interviews for the purpose. Attitude or opinion research aims at discovering the underlying motives and desires, using in depth interviews for the purpose. Other techniques of such research are word association test, sentence completion test, story completion test and similar other projective techniques. Attitude or opinion research i.e. research designed to find out how people feel or what they thought about a particular subject or institution is also qualitative research. Qualitative research is especially important in the behavioural sciences where the aim is to discover the underlying motives of human behaviour. Through such research we can analyse the various factors which motivate people to behave in a particular manner or which make people like or dislike a particular thing. It may be stated, however, that to apply for qualitative research in practice is relatively a difficult job and therefore, while doing such research, one should seek guidance from experimental psychologists.

3.5 Conceptual vs. Empirical Research

Conceptual research is that related to some abstract idea(s) or theory. It is generally used by philosophers and thinkers to develop new concepts or to reinterpret existing ones. On the other hand, empirical research relies on experience or observation alone, often without due regard for system and theory. It is data-based research, coming up with conclusions which are capable of being verified by observation or experiment. We can also call it as experimental type of research. In such a research, the researcher must first provide himself with a working hypothesis or guess as to the probable results. He then works to get enough facts (data) to prove or disprove his hypothesis. He then sets up experimental designs which he thinks will manipulate the persons or the materials concerned so as to bring forth the desired information. Such research is thus characterised by the experimenter’s control over the variables under study and his deliberate manipulation of one of them to study its effects. Empirical research is appropriate when proof is sought that certain variables affect other variables in some way. Evidence gathered through experiments or empirical studies are considered to be the most powerful support possible for testing a given hypothesis.

3.6 Longitudinal Research

All other types of research are variations of one or more of the above stated approaches, based on either the purpose of research, or the time required to accomplish research, on the environment in which research is done, or on the basis of some other similar factors. From the point of view of time, we can think of research either as ***one-time research*** **or** ***longitudinal research***. In the former case the research is confined to a single time-period, whereas in the latter case the research is carried on over several time-periods. Research can be *field-setting research or laboratory research or simulation research*, depending upon the environment in which it is to be carried out.

Research can as well be understood as *clinical or diagnostic research.* Such research studies usually go deep into the causes of things or events that interest us, using very small samples and very deep probing data gathering devices. The research may be *exploratory* or it may be formalised. The objective of exploratory research studies is the development of hypotheses rather than their testing, whereas formalised research studies are those with substantial structure and with specific hypotheses to be tested. *Historical research* is that which utilises historical sources like documents, remains, etc. to study events or ideas of the past, including the philosophy of persons and groups at any remote point of time.

Research can also be classified as *conclusion-oriented* and *decision-oriented*. While doing conclusion-oriented research, a researcher is free to pick up a problem, redesign the enquiry as he proceeds and is prepared to conceptualise as he wishes. Decision-oriented research is always for the need of a decision maker and the researcher in this case is not free to embark upon research according to his own inclination. Operations research is an example of decision oriented research since it is a scientific method of providing executive departments with a quantitative basis for decisions regarding operations under their control.

3.7 Cross-sectional Studies (Non-experimental)

It is also one time or at a point of time study of all persons in a representative sample of a specific population such as examination of all children in age group 5 to 14 years, detection of cancer cases and study of factors that lead to cancer, examination of children in age group 0 to 6 years for classifying into nutritional grades, finding prevalence of pregnancy in age group 20 to 30 years of married women or morbidity due to cancer, paralytic polio, etc. Such studies indicate point prevalence, i.e. number of cases at a time of study. Field surveys in health or disease problems in a community and census are other examples of cross sectional studies or research as the case may be or addressed.

3.8 Cohort Studies

Is a group of persons exposed to same sort of environment such as newborns, women between 15 and 45 years of age, or workers exposed to radiation or other kinds of hazards in occupation? Cohort study could be prospective such as follow-up of morbidity and mortality in infants from births to one year of age or it could be retrospective inquiry such as number of person in the same population who suffered from typhoid in last five years.

3.9 Control Studies

Most of the experimental studies almost always need a control as a yard stick of evidence, very rarely it may not be required as for trial in a fatal disease like rabies. It may be unethical to withhold an established treatment to control cases in which life is at stake or there is fear of serious after effect. It would not be fair to withhold antibiotics in control cases of typhoid or lobar pneumonia:

4.0 CONCLUSION

Research in common parlance refers to a search for knowledge. When the unknown confronts us, more and more our inquisitiveness makes us probe and attain understanding of the unknown. This inquisitiveness is the mother of all knowledge and the method, which one employs for obtaining the knowledge of whatever the unknown, can be termed as research. Research is an academic activity and as such the term should be used in a technical sense. Research comprises redefining problems, formulating hypothesis or suggested solution; and at last carefully testing the conclusion to determine whether they fit the formulating hypothesis. Research is an original contribution to the existing stock of knowledge through objective and systematic method of finding solution to a problem is research. Various types of researches include, Descriptive, Analytic Applied, Fundamental, Quantitative, Qualitative, Conceptual, Empirical research Longitudinal, Cross-sectional, Cohort, and Control Studies.

5.0 SUMMARY

In this unit, you have learnt the following;

- The meaning of the word “research**”**
- The meaning of descriptive and analytic research.
- Applied vs. Fundamental research
- Quantitative vs. Qualitative research

- Conceptual vs. Empirical research
- Longitudinal research
- Cross-sectional research
- Cohort study/research

6.0 TUTOR-MARKED ASSIGNMENT

1. Concisely explain the meaning of research.
2. Clearly distinguish between longitudinal and Cross-sectional Researches.
3. Write briefly on the following
   a) Descriptive research.
   b) Analytic research.
   c) Applied research.
   d) Fundamental research
   e) Quantitative research
   f) Qualitative research
   g) Conceptual research
   h) Empirical research
   i) Longitudinal research
   j) Cross-sectional research
   k) Cohort study/research

7.0 REFERENCES/FURTHER READING

Bushit, D.B.&Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nded.).Jaypee Brothers Medical Publishers Ltd. New Delhi:ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New AgeInternationalPublishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6thed.).Jaypee Brothers MedicalPublishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7thed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

UNIT 2 OTHER RESEARCH FEATURES AND CHARACTERISTICS

CONTENTS

1.0 INTRODUCTION

This involves, Approaches to Research, Significance of Research, Research Methods versus Research Methodology, Research Problem and Research Design. There are two basic approaches to research, viz., *quantitative approach* and the *qualitative approach.* The former involves the generation of data in quantitative form which can be subjected to rigorous quantitative analysis in a formal and rigid fashion. We shall consider this into details later in the course of this unit.

Research significantly impact Economic, Social, Political, Medical, and Scientific spheres of human endeavour, again we shall see how shortly. Research has methods of operation and also methodology, the distinction between these two important terminologies as far as research is concerned shall be explicitly X-rayed as we move along in this unit. Equally worth mentioning in this aspect are the definitions and implications of research problem and research design. We shall soon consider these features of research in the later part of this unit.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- list the different approaches to research
- explain the significance of research
- discuss research objective
- explain research methods versus methodology
- discuss research problem.

3.0 MAIN CONTENT

3.1 Research Approaches

The above description of the type's research brings to light the fact that there are two basic approaches to research, viz., *quantitative approach* and the *qualitative approach.* The former involves the generation of data in quantitative form which can be subjected to rigorous quantitative analysis in a formal and rigid fashion. This approach can be further sub-classified into *inferential, experimental* and *simulation* approaches to research. The purpose of *inferential* approach is to form a data base to infer characteristics or relationships of population. This usually means survey research where a sample of population is studied (questioned or observed) to determine its characteristics and it is then inferred that the population has the same characteristics. *Experimental approach* is characterised by much greater control over the research environment and in this case some variables are manipulated to observe their effect on other variables. *Simulation approach* involves the construction of an artificial environment within which relevant information and data can be generated. This permits an observation of the dynamic behaviour of a system (or its sub-system) under controlled conditions. The term 'simulation' in the context of business and social sciences applications refers to "the operation of a numerical model that represents the structure of a dynamic process. Given the values of initial conditions, parameters and exogenous variables, a simulation is run to represent the behaviour of the process over time". Simulation approach can also be useful in building models for understanding future conditions.

Qualitative approach to research is concerned with subjective assessment of attitudes, opinions and behaviour. Research in such a situation is a function of researcher's insights and impressions. Such an approach to research generates results either in non-quantitative form or in the form which is not subjected to rigorous quantitative and analysis. Generally, the techniques of focus group interviews, projective techniques and depth interviews are used. All these are explained at length in chapters that follow.

3.2 Significance of Research

- The role of research in several fields of applied economics whether related to business or to the economy as a whole has greatly increased in modern times.
- Research provides the basis for nearly all government policies in our economic system.
- Research has its special significance in solving various operational and planning problems of business and industry.
- Research is equally important for social scientists in studying social relationships and in seeking answers to various social problems.

In addition to what has been stated above, the significance of research can also be understood keeping in view the following points:

a. To those students who are to write a master's or Ph.D. thesis, research may mean a careerism or a way to attain a high position in the social structure;
b. To professionals in research methodology, research may mean a source of livelihood.
c. To philosophers and thinkers research may mean the outlet for new ideas and insights;
d. To literary men and women research may mean the development of new styles and creative work; and
e. To analysis and intellectuals research may mean the development of new theories.

3.3 Objectives of Research

The purpose of research is to discover answers to questions through the application of scientific procedures. The main aim of research is to find out the truth which is hidden and which has not been discovered yet. Though each research study has its own specific purpose, we mention some general objectives of research below:

– To gain familiarity with a phenomenon or to achieve new insights into it
– To portray accurately the characteristics of a particular individual, situation or group
– To determine the frequency with which something occurs or with which it is associated with something else.
– To test a hypothesis of a causal relationship between variables.

3.4 Research Methods versus Methodology

It seems appropriate at this juncture to explain the difference between research methods and methodology. Research methods may be understood as all those methods / techniques that are used for conduction of research. Research methods or techniques that refer to the methods the researchers use in performing research operations. In other words, all those methods which are used by the researcher during the course of studying his research problem are termed as research methods. Since the object of research, particularly the applied research, is to arrive at a solution to a given problem, the available data and the unknown aspects of the problem have to be related to each other to make a solution possible. Keeping this in view, research methods can be put into the following three groups

In the first group we include those methods which are concerned with the collection of data. These methods will be used where the data already available is not sufficient to arrive at the required solution.

The second group consists of those statistical techniques which are used for establishing relationships between the data and the unknown.

The third group consists of those methods which are used to evaluate the accuracy of the results obtained.

Research methods falling in the above stated last two groups are generally taken as the analytical tools of research.

At times, a distinction is also made between research methods and techniques. Research techniques refer to the behaviour and instruments we use in performing research operations such as making observations, and recording data, techniques of processing data and the like. Research methods refer to the behaviour and instrument used in selecting and constructing research technique. For instance, the difference between methods and techniques of data collection can be better understood from the details given in the following chart:

Table 2.1: Research Methodology

Type	Methods	Technique
1 Library Research	i) Analysis of historical records ii) Analysis of Documents	Recoding of notes, Content analysis, Tape and Film listening and analysis Statistical compilation and manipulation, reference and abstract guides, content analysis.
2 Field Research	i) Non –participant direct observation ii)Participant observation iii)Mass observation iv)Mail questionnaire v) Opinionnaire vi)Personal interview vii)Focused interview viii)Group interview ix) Telephone survey x) Case study and life history	Observational behavioural scales, use of cards, etc. Interactional recording, possible use of tape recorders, photographic techniques Recording mass behaviour, interview using independent observers in public places. Identification of social and economic background of respondents. Use of attitude scales , projective techniques, use of sociometric scales Interviewer uses a detailed schedule with open and closed ended questions. Interviewer focuses attention upon a given experience and its effects. Small group of respondents are interviewed simultaneously. Use as a survey technique for information and for discerning opinion, may also be used as a follow up of questionnaire. Cross-sectional collection of data for intensive analysis, longitudinal collection of data of intensive character.
3)Laboratory Research	Small group study of random behaviour, play and role analysis	Use of audio visual recording devices, use of observers, etc.

From what has been stated above, we can say that methods are more general. It is the method that generates techniques. However, in practice, the two terms are taken as interchangeable and when we talk of research methods we do, by implication, include research techniques within their compass.

Research Methodology is a way to systematically solve the research problem. It may be understood as a science of studying how research is done scientifically. In it we study the various steps that are generally adopted by a researcher in studying his research problem along with the logic behind them. It is necessary for researcher to know not only the research methods / techniques but also the methodology. Researcher not only need to know how to develop certain indices or tests, how to calculate the mean, the mode, the median or the standard deviation or chi square, how to apply particular research techniques, but they also need to know which of these methods or techniques are relevant and which are not, and what they would mean and indicate.

Researchers also need to understand the assumption underlying various techniques and they need to know the criteria by which they can decide that certain techniques and procedures will be applicable to certain problems and others will not: All this means that it is necessary for the researchers to design a methodology for his problem as the same may differ from problem to problem. For example, an architect, who designs a building, has to consciously evaluate the basis of his decisions, i.e., he has to evaluate why and on what basis he selects particular size, number and location of doors, windows and ventilators, uses particular materials and not others and the like. Similarly, in research the scientist has to expose the research decisions to evaluation before they are implemented. He has to specify very clearly and precisely what decisions he selects and why he select them so that they can be evaluated by others also.

From what has been stated above, we can say that research methodology has many dimensions and research do constitute a part of the research methodology. The scope of research methodology is wider than research methods. *Thus, when we talk about research methodology we not only talk about research methods but also the logic behind the methods we use in the context of our research study and explain why we are using a particular technique and why we are not using other so that research results are capable of being evaluated either by the researcher or by others*. Why a researcher study has been undertaken, how the research has been defined and why the hypothesis has been formulated, what data have been collected and what particular method has been adopted, why particular technique of analysing data has been used are a host of similar questions that are answered when we talk about research methodology concerning a research problem or study.

3.5 Research and Scientific Method

For a clear perception of the term research, one should know the meaning of scientific method. The two terms, research and scientific methods, are closely related. In research, we study the nature, reasons, and the consequences of a set of circumstances which are controlled experimentally or observed as they appear. Researcher is usually interested in particular results, the repetitions of those results, and generalisations. On the other hand, the philosophy common to all research methods and techniques, although they may vary considerably from one science to another is usually given the name of scientific method. In this context, Karl Pearson writes, "The scientific method is one and the same in the branches (of science) and that method is the method of all logically trained minds…the unity of all sciences consists alone in its methods, not in its material: the man who classifies facts of any kind whatever, who sees their mutual relation and describes their consequences, is applying the scientific method and is a man of science." The scientific method is the pursuit of truth as determined by logical by logical considerations. The ideal of science is to achieve a systematic interrelation of facts. Scientific method attempts to achieve this ideal by experimentation, observation, logical arguments from accepted postulates and a combination of these three in varying proportions. In scientific method, logic aids in their formulating propositions explicitly and accurately so that their possible alternatives become clear. Further, logic develops the consequences of such alternatives, and when these are compared with observable phenomena, it becomes possible for the researcher or the scientist to state which alternative is the most in harmony with the observed facts. All this is done through experimentation and survey investigations which constitute the integral parts of scientific method.

3.6 Defining the Research Problem

In research process, the first and foremost step happens to be that of selecting and properly defining a research problem. A researcher must find the problem and formulate it so that it becomes susceptible to research. Like a medical doctor, a researcher must examine all the symptoms (presented to his or observed by his) concerning a problem before he can diagnose correctly. To define a problem correctly, a researcher must know: what a problem is?

3.6.1 What is a Research Problem?

A research problem, in general, refers to some difficulty which a research experiences in the context of either a theoretical or practical situation and wants to obtain a solution for the same. Usually we say that a research problem does exist if the following conditions are met with:

(i) there must be an individual (or a group or an organisation), let us call it 'I,' to whom the problem can be attributed the individual or the organisation, as the case may be, occupies an environment, say 'N,' which is defined by values of the uncontrolled variables Y_I

There must be at least two courses of action, say C1 andC2 to be pursued. A course of action is defined by one or more values of the controlled variables. For example, the number of items purchased at a specified time is said to be one course of action.

There must be at least two possible outcomes, say O1 andO2 of a course of action, of which one should be preferable to the other. In other words, this means that there must be at least one outcome that the researcher wants, i.e. an objective.

The courses of action available must provide some chance of obtaining the objective, but they cannot provide the same chance, otherwise the choice would not matter. Thus, if P ($O_{J,}$ \ I, C, N) represents the probability that an outcome O_J will occur, if I selects CJ in N, then P ($O_{J,}$ \ *I,* C_1, N) ≠ P ($O_{J,}$ \ *I*, C_2, N). In simple words, we can say that the choices must have unequal efficiencies for the desired outcomes.

Over the above, these conditions, the individual or the individual or the organisation can be said to have the problem if *I* does not know what course of action is best i.e., *I*, must be in doubt about the solution. Thus, an individual or group of persons can be said to have a problem which can be technically described as a research problem, if they (individual or the group), having one or more desired outcomes, are confronted with two or more courses of action that have some but not equal efficiency for the desired objective(s) and are in doubt about which course of action is best.

We can thus state the components of research problem as follows;

i. There must be some individual or group which have some problem.

ii. There must be some objective(s) to be attained. If one wants nothing, one cannot have a problem.
iii. From the popular Yoruba adage," a Dog does not bark without a cause"
iv. There must be alternatives means (or causes of action) for obtaining the objective(s) one wishes to attain. This means that there must be at least two means available to a researcher for if he has no choice of means, he cannot have a problem.
v. There must remain some doubt in the mind of a researcher with regard to the selection of alternatives. This means that the researcher must answer the question concerning the relative efficiency of the possible alternatives.
vi. There must be some environment(s) to which the difficulty pertains.

Thus a research problem is one which requires a researcher to find out the best solution for a given problem, i.e. to find out by which course of action the objective can be attained optimally in the context of a given environment. There are several factors which may result in making the problem complicated. for instance, the environment may change affecting the efficiencies of the courses of action or the values of the outcomes; the number of alternatives courses of action may be very large; persons not involved in making decisions may be affected by it and react to it favourably or unfavourably, and similar other factors. All such elements (or at least the important ones) may be thought of in the context of a research problem.

3.6.2 Selecting the Problem

The research problem undertaken for study must be carefully selected. The task is a difficult one, although it may not appear to be so. Help may be taken from a research guide in this connection. Nevertheless, every researcher must find out his own salvation for research problem cannot be borrowed. A problem must spring from the research's mind like a plant springing from its own seed. If our eyes need glasses, it is not the optician alone who decides about the number of the lens cooperating with him. Thus, a research guide can at the most only help a researcher choose a subject. However, the following points may be observed by a researcher in selecting a research problem or a subject for research:

I. Subject which is overdone should not be normally chosen, for it will be a difficult task to throw any new light in such a case.
II. Controversial subject should not become the choice of an average researcher.

III. Too narrow or too vague problem should be avoided.

IV. The subject selected for research should be familiar and feasible so that the related research material or sources of research are within one's reach. Even then it is quite difficult to supply definitive ideas concerning how a researcher should obtain ideas for his research. For this purpose, a researcher should contact an expert or a professor in the university who is already engaged in research. He may as well read articles published in current literature available on the subject and may think how techniques and ideas discussed there in might be applied to the solution of other problems. He may discuss with others what he has in mind concerning a problem. In this way he should make all possible efforts in selecting a problem.

V. The importance of the subject, the qualifications and the training of a researcher, the costs involved and the time factor are few other criteria that must also be considered in selecting a problem. In other words, before the final selection of a problem is done, a researcher must ask himself the following questions:

   a. Whether he is well equipped in terms of his background to carry out the research?
   b. Whether the study falls within the budget he can afford?
   c. Whether the necessary cooperation can be obtained from those who must participate in research as subjects?

If the answers to all these questions are in the affirmative, one may become sure so far as the practicability of the study is concerned.

The selection of a problem must be preceded by a preliminary study. This may not be necessary when the problem requires the conduct of a research closely similar to one that has already been done. But when the field of inquiry is relatively new and does not have available a set of well-developed techniques, a brief feasibility study must always be undertaken.

If the subject for research is selected properly by observing the above mentioned points, the research will not be a boring drudgery, rather it will be love's labour. In fact, zest for work is a must. The subject or the problem selected must involve the researcher and must have an upper most place in his mind so that he may undertake all pains needed for the study.

3.6.3 Need of Defining the Problem

Quite often we all hear that a problem clearly stated is a problem half solved. This statement signifies the need for defining a research problem.

The problem to be investigated must be defined unambiguously for that will help to discriminate relevant data from the irrelevant ones. A proper definition of research problem will enable the researcher to be on the track whereas an ill-defined problem may create hurdles. Questions like: what data is to be collected? What characteristics of data are irrelevant and need to be studied? What relations are to be explored? What techniques are to be used for the purpose? And similar other questions crop up in the mind of the researcher who can well plan his strategy and find answers to all such question crop up in the mind of the researcher who can well plan his strategy and find answers to all such questions only when the research problem has been well defined. Thus, defining a research problem properly is a prerequisite for any study and is a step of the highest importance. In fact, formulation of a problem is often more essential than its solution. It is only on the careful detailing of the research problem that we can work out the research design and can smoothly carry on all the consequential steps involved while doing research. Experimentation is done to test hypotheses and to discover new relationships, if any, among variables. However, sometimes the conclusions drawn on the basis of experimental data may be misleading for faulty assumptions, poorly designed experiments, badly executed experiments or faulty interpretations. As such the researcher must pay all possible attention while developing the experimental design and drawing inferences. The purpose of survey investigations may also be to provide scientifically gathered information to work as a basis for the researchers for their conclusions.

The scientific method is, thus, based on certain basic postulates which can be stated as under:

1. It relies on empirical evidence;
2. It utilises relevant concepts;
3. It is committed to only objective considerations;
4. It aims at nothing but making only adequate and correct statements about population objects;
5. It results into probabilistic predictions;
6. Its methodology is made known to all concerned for critical scrutiny and are for use in testing the conclusions through replication;
7. It aims at formulating most general axioms or what can be termed as scientific theories.

Thus, the scientific method encourages a rigorous, method wherein the researcher is guided by the rules of logical reasoning, a method wherein the investigation proceeds in an orderly manner and a method that implies internal consistency.

3.6.4 Technique Involved in Defining a Problem

Defining research problem properly and clearly is a crucial part of a research study and must in no case be accomplished hurriedly. However, in practice this is frequently overlooked which causes a lot of problems later on. Hence, the research problem should be defined in a systematic manner, giving due consideration to all relating points. The technique for the purpose involves the undertaking of the following steps generally one after the other:

i. Statement of the problem in a general way;
ii. Understanding the nature of the problem;
iii. Surveying the available literature;
iv. Developing the ideas through discussions; and
v. Rephrasing the research problem into a working proposition.

In addition to what has been stated above, the following points must also be observed while defining a research problem:

(a) Technical terms and words or phrases, with special meanings used in the statement of the problem, should be clearly defined.
(b) Basic assumptions or postulates (if any) relating problem should be clearly stated.
(c) A straight forward statement of the value of the investigation (i.e., the criteria for the selection of the problem) should be provided.
(d) The suitability of the time-period and the sources of data available must also be considered by the researcher in defining the problem.
(e) The scope of the investigation or the limits within which the problem is to be studied must be mentioned explicitly in defining a research problem.

3.7 Research Design

The formidable problem that follows the defining the research problem is the preparation of arrangement of condition for collection and analysis of data in a manner that aims to combine the relevance to the research purpose with economy in procedure. In fact, the research design is the conceptual structure within which is research is conducted; it constitutes the blueprint for the collection, measurement and analysis of data. As such the design includes an outline of what the researcher will do from writing the hypothesis and its operational implication to the final analysis of data. More explicitly, the design decisions happen to be in respect of:

I. What is the study about?
II. Why is the study being made?
III. Where will the study be carried out?
IV. What type of data is required?
V. Where can the required data be found?
VI. What period of time will the study include?
VII. What will be the sample design?
VIII. What techniques of data collection will be used?
IX. How will the data be analysed?
X. In what style will the report be prepared?

Keeping in view above stated design decisions, one may split the overall research design into the following parts:

(a) *The sampling design* which deals with the method of selecting items to be observed for the given study.
(b) *The observational design* which relates to the condition under which the observation are to be made
(c) *The statistical design* which concerns the question of how many items are to be observed and how the information and data gathered is to be analysed.
(d) The *operational design* who deals with the techniques which procedures specified in the sampling, statistical and observational designs can be carried out.

From what has been stated above, we can state the important features of a research design as under.

It is plan that specifies the sources and types of information relevant to the research problem.

Keeping in view the above stated design decisions, one may split the overall research design into the following part.

(a) the sampling design which deals with the method of selecting items to be observed for the given study.
(b) the observational design which relates to the conditions under which the observations are to be made.
(c) the statistical design which concerns the question of how many items are to be observed and how the information and data gathered is to be analysed.

(d) the operational design which deals with the techniques by which the procedures specified in the sampling, statistical and observational designs can be carried out.

From what has been stated above, we can state the importance features of a research design as under

(i) It is a plant that specifies the sources and types of information relevant to the research problem.
(ii) It is a strategy specifying which approach will be used for gathering and analysing the data.
(iii) It also includes the time and cost budgets since most studies are done under these two constraints.

In brief, research designs must, at least, contain

(a) a clear statement of the research problem;
(b) procedures and techniques to be used for gathering information;
(c) the population to be studied;
(d) methods to be used in processing and analysing data.

3.7.1 Need for Research Design

Research design is needed because it facilitates the smooth sailing of the various research operations, thereby making research as efficient as possible yielding maximal information with minimal expenditure of effort, time and money. Just as for better, economical and attractive construction of a house, we need a blue print (or what is called the map of the house) well thought out and prepared by an expert architect, similarly, we need research design or a plan in advance of data collection and analysis for our research project. Research design stands for advance planning of the method adopted for collecting the relevant data and the techniques to be used in their analysis, keeping in view the objective of the research and the availability of staff, time and money. Preparation of the research design should be done with great care as any error in it may upset the entire project. Research design, in fact, has a great bearing on the reliability of the results arrived at and as such constitutes the firm foundation of the entire edifice of the research work.

Even then the need for a well thought our research design is at times not realised by many. The importance which this problem deserves is not given to it. As a result many researches do not serve the purpose for which they are undertaken. In fact, they may even give misleading conclusions.

Thoughtlessness in designing may results in rendering the research exercise futile. It is, therefore, imperative that an efficient and appropriate design must be prepared before starting research operations. The design helps the research to organise his ideal in a form whereby it will be possible for him to look for flaws and in adequacies. Such a design can even be given to other for their comment and critical evaluation. In the absence of such a course of action, it will be difficult for the critic to provide a comprehensive review of the proposed study.

3.7.2 Features of a Good Design

A good design is often characterised by adjectives like flexible, appropriate, efficient, and economical and so on. Generally, the design which minimises bias and maximises the reliability of the data collected and analysed is considered a good design. The design which gives the smallest experimental error is supposed to be the best design in many investigations. Similarly, a design which yields maximal information and provides an opportunity for considering many aspects of a problem is considered most appropriate and efficient design in respect of many research problems. Thus, the question of good design is related to the purpose or objectives of the research problem and also with the nature of the problem studied. A design may be quite suitable in one case, but may be found wanting in one respect or the other in the context of some other research problem. One single design cannot serve the purpose of all types of research problems.

A research design appropriate for a particular research problem, usually involve the consideration of the following factors;

I. The means of obtaining information;
II. The availability and skills of the research and his staff, if any;
III. The objective of the problem studied;
IV. The nature of the problem to be studied; and
V. The availability of time and money for the research work.

If the research study happens to be an exploratory or a formulative one, where on the major emphasis is on discovery of ideas and insights, the research design most appropriate must be flexible enough to permit the consideration of many different aspect of a phenomenon. But when the purpose of a study is accurate description or an association between variables (or in what are called the descriptive studies), accuracy becomes a major consideration and research design which minimises bias and maximises the reliability of the evidence collected is considered a good design. Studies involving the testing of a hypothesis of a causal relationship between variables require a design which will permit interference about causality in addition to the minimisation of bias and maximisation of reliability. But in practice it is most difficult task to put a particular study in a particular group, for a given research may have two or more of the function of different studies. It is only on the basis of its primary function that a study becategorised either as an exploratory or descriptive or hypothesis-testing study and accordingly the choice of a research design may be made in a particular study. Besides, money, time, skill of the research staff and the means of obtaining the information must be given due weightage while working out the relevant details of the research design, survey design, sample design and the like.

3.7.3 Important Concepts Relating to Research Design

Before describing the different research designs, it will be appropriate to explain the various concepts relating to design so that these may be better and easily understood.

3.8 Dependent and Independent Variables

A concept which can take on different quantitative value is called a *variable.* As such the concepts like weight, height; income are all example of variables. Quantitative phenomena (or the attributes) are also quantified on the basis of the presence or absence of the concerning attributes phenomena may take on quantitatively different values even in decimal points are called "continuous variable". But all variables are not continuous. If they can be expressed in integral values, they are non-continuous variable or in statistical language "discrete variables". Age is an example of continuous variable, but the number of children is an example of non- continuous variable. If one variable depends upon or is a consequence of the other variable, it is termed as dependent variable, and the variable that is antecedent to the dependent variable is termed as independent variable. For instance, if we say that height depends upon age, then height is a dependent variable and age is an independent variable.

Further, if in addition to being dependent upon age, height also depends on individual sex, then height is a dependent variable and age and sex are independent variables. Similarly, ready made films and lectures are example of independent variables, whereas behavioral changes, occurring as a result of the environmental manipulations, are example of dependent variables.

4.0 CONCLUSION

Research features and characteristics involve, approaches to research, significance of research, research methods versus research methodology, research problem and research design. *Quantitative approach* and the *qualitative approach* are the two basic approaches to research. The former involves the generation of data in quantitative form which can be subjected to rigorous quantitative analysis in a formal and rigid fashion. Research significantly impacts Economic, Social, Political, Medical, and Scientific spheres of human endeavour.

5.0 SUMMARY

In this unit, you learnt the following basic features of research**:**

- Approaches to research
- Significance of research
- Research Objective
- Research Methods versus Methodology
- Research Problem
- Research Design

6.0 TUTOR-MARKED ASSIGNMENT

1 What are the approaches to research?
2 Distinguish as far as you can, research method and methodology of research.
3 Write concise essays on the following;
   i) Research Objective
   ii) Research Problem
   iii) Research Design
4 What are independent variables as opposed to dependent variables in the fields of research?

7.0 REFERENCES/FURTHER READING

Bushit, D.B. & Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2^{nd} ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim&Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

UNIT 3 SAMPLING

CONTENTS

1.0 INTRODUCTION

It is not possible to include each member (sampling unit) of the population in an experimental study or enquiry or examine all the millions people in Nigeria to find the prevalence of tuberculosis or the efficacy of a drug in all the patient suffering from a particular disease. A daily life example is that of cooking beans. A housewife just pick *a few grains* of beans from a cooking vessel and gets a fairly good idea *whether* the entire lot of beans is fully cooked or it requires more cooking. Further, covering the entire population may be less accurate because a large number of investigators is required to complete a huge task. Their uniformity and correctness may vary, collection will be costly, time consuming and laborious. Because of all such difficulties, we prefer to use an appropriate sampling technique.

In medical studies, the sampling data are collected from a population or universe sufficiently large and representative of the population study, chosen by standard sampling technique.

The population or universe must clearly be defined before drawing sample. For example, population may be an entire group of defined people such as doctors, all members of household, women 15 to 49 years of age and so on, about whom information is required. It has to be further which doctors, those in services or in general practice, which women, married or unmarried, etc.

A value calculated from a defined population, such as mean (μ), standard deviation (δ), or standard error of mean (×) is called a *parameter*. It is a constant value because it covers all the member of the population. A value calculated from a sample is called a *statistics* such as mean (×), standard deviation (s) and proportion (p).

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- generalised sampling characteristics in research planning and implementation
- explain the meaning of precision (sample size and its determination)
- discuss characteristics during sampling
- differentiate sampling procedures.

3.0 MAIN CONTENT

3.1 Sample Characteristics

Sample is any part of the population. Large number of samples may be taken from the same population. Still all members may not be covered. The composition of sample may vary in size, quality, and techniques in drawing the sample thereby their statistics also vary. Inference drawn from a sample refers to the defined population (universe) from which sample or samples are drawn and not to any other population.

Such inference or conclusion drawn from the samples applied to the whole population or universe but generalisations are valid, only if the sample is sufficiently large and unbiased i.e. representative of the entire population from which it is drawn. A representative sample will have its statistics almost equal to the parameters of the entire population. There will still be a chance difference or error of chance difference or error of chance, which can be calculated from the representative sample. The difference can be reduced but not eliminated.

There are two main characteristics of a representative sample;

1. Precision which implies the size of the sample
2. Unbiased character

These two qualities help to fulfil the objectives of sample stated above. Then we can decide, whether the difference observed in values of samples are due to chance or some other factors when compared with population parameters or statistics of another sample.

3.2 Precision

Precision depends on the sample size this implies the size, i.e. numbers in the samples depending on the purpose ordinarily it should not be less than 30. A sample, small in size, is a biased one and never be depended upon for drawing any conclusions. In typhoid or lobar pneumonia, the case mortality rate or probability of dying after any treatment varied from 20 per cent to 30per cent before drugs like antibiotics and sulpha were discovered. If a highly qualified doctor, by chance treats only three cases, out of those 30per cent which are going to die, he gets 100per cent mortality and might be considered inefficient. Another doctor to whom three cases coming for treatment from the remaining 70per cent will be considered more efficient and may even be eulogized. As a matter of fact, neither doctor is to be blamed nor accredited. The wrong conclusion is due to the small size of samples by applying random technique, mortality in both cases would have been 25-30per cent.

While comparing prevalence of diabetes in different occupations in well conducted diabetes survey the investigator concluded the prevalence of diabetes amongst soldiers was 66per cent, when only threesoldiers out of which two had diabetes, while entered the study of a mixed population of 5000. He compared diabetes prevalence in soldiers with that in traders who were in large number in the sample. Thus the sample is very vital in any scientific study. Therefore, how large a sample is considered large enough? Normally, the cut off is taken at 30. A sample of size greater than 30 is considered large enough for statistical purposes.

If samples are small in size such as 2, 4, 6, 20, etc., their ‘means’ will be quite different from each other as well as from the population mean, and when the size is 30 or more their means will be closer to each other and would be in the neighbourhood of population mean. In other words, a small sample lacks precision.

Precision is measured by the formula; Precision = $\sqrt{n}/s$'s' is the Standard Deviation 'SD' of a sample in the estimate of 'σ' which is the 'SD' of the population mean. So nearness of s to σ i.e., the precision will be directly proportional to the square root of the sample size 'n' e.g. if the n is increased four times, the precision will be doubled.

Examples
If s is 2 and n is 4, the precision $\sqrt{n}/s = \sqrt{4}/2 = 2/2 = 1$
Assuming that s = 2 and varying the size of n, we calculate the precision as shown below;
If n= 16, precision = $\sqrt{16}/2 = 4/2 = 2$(Precision becomes doubled if sample size is increased to four times).
If n is 36, precision = $\sqrt{16}/2 = 6/2 = 3$(Precision is trebled if n is increased to 9 times).
If n is 64 precision = $\sqrt{64}/2 = 8/2 = 4$ (Precision is quadrupled if n is increased to 16 times and so on).
In case of an experiment, if the results are not decisive about the variation by chance or due to an external factor, increase the size of sample. e.g. if calculated probability(p) is round about 0.05 the conclusion can be confirmed by increasing the size of the sample. To choose the suitable number in order to reduce the sampling error to the minimum and to clear the concept of determining the suitable size of sample to the reader, some of the methods are recommended here.

For Quantitative Data
In such data, we deal with the means of a sample and of the universe. If the SD (σ) in the population is known from the past experience, the size of sample can be determined by the following formulae by the desirable *allowable error* (L). At five per cent risk the true estimate will lie beyond the allowable error (variation).

Hence, the first step is to decide how large an error due to sampling defects can be allowed or tolerated in the estimates. Such allowable error has to be stated by the investigator.

The second step is to express the allowable error in terms of confidence limits. At Suppose L is the allowable error in the sample mean and we are willing to take a five per cent chance that the error will exceed L. so we may put:
$L=2\sigma/\sqrt{n}$ or $\sqrt{n}=2\sigma/L$ or $n=4\sigma^2/L^2$
If $L=1$ and $\sigma^2=25$, $n= 4*25/1^2=100$
In such cases, the investigation may start which an assumed SD the allowable error specified by the experimenter. In case SD is not known,

preliminary investigation or a pilot survey may have to be carried out to estimate the population SD.

Examples

1. Mean pulse rate of a population is believed to be 70 per minute with a standard deviation of 8 beats. Calculate the minimum size of the sample to verify this, if allowable error L=$\pm$ 1 beat at five per cent risk.
   $n=4\sigma^2/L^2$=4*8*8/1*1=256
   (Source: Snedecor GW, fifth edition p.502)
   If L=+ 2 beats with 5% risk
   N= 4*8*8/2*2=64
   If L is less, n will be more, i.e., larger the size, lesser will be the error.
2. Mean systolic blood pressure in one college students was found to be 120 with SD of 10. Calculate the minimum size of the sample to verify the result if allowable error is $\pm$ 2 at 5% risk.
   n= $4\sigma^2/L^2$=4*10*10/2*2=100
   For quantitative Data

In such data, we deal with proportion such as morbidity rates and cure rates. For finding the suitable size of the sample, the assumption usually made is that the allowable error does not exceed 10per cent or 20per cent of the positive character.

The size can be calculated by the following formula with a desired allowable error (L) at five per cent risk that the true estimate will not exceed allowable error by 10per cent or 20per cent of 'p'

$$n=4pq/L^2$$

where 'P' is the positive character, q=1-P and L=allowable error, 10per cent or 20per cent of 'P'.

Examples

1. Incidence rate in the last influenza epidemic was found to be 50 per thousand (5%) of the population exposed. What should be the size of sample to find incidence rate in the current epidemic if allowable error is 10per cent and 20per cent?
   n= $4pq/L^2$, p=5%, q=95%
   if L=10% of p=5*10/100=0.5%
   $n=4pq/L^2$=4*5*95/0.5*0.5=7600
   If L=20% of p=5*20/100=1%,
   $n=4pq/L^2$=4*5*95/1*1=1900

   The larger the permissible error, the smaller will be the size of sample required for both types of data.

2. Hookworm prevalence rate was 30per cent before the specific treatment and adoption of others measures. Calculate the size of the sample required to find the prevalence rate now if allowable error is 10per cent and 20per cent.
   If L=10% of p=30*10/100=3

At 5% risk, n= $4pq/L^2$ = 4x30x70/3x3= 933.3(or 934 to round off the size.
If L = 20% of P = 30 x20 / 100 = 6
At 5% risk n = $4pq/L^2$ = 4x30 x70/6x6 =233.3(or 234 to round off the size)

Thus if we allow a small error, the required sample size will be much larger as compared to one when the allowable error is increased.

3.3 Unbiased Character

Sample bias: Bias comes in when the samples from a population are not chosen at random or samples are not drawn from similar populations. Bias may creep in due to non-sampling errors. Bias due to sampling defects alone is discussed here. When there is bias, the statistics of samples like mean ($\underline{x}$) will be away from the population parameter (μ).

The mean weight of 50 babies at birth would be nearly the mean weight of the total babies born in the population. If these babies were born in a private maternity home where only women from the affluent families come for delivery, then, they form a biased or selected sample of all babies. They may give an unbiased estimate of mean baby weight in well-to-do class but would seriously overestimate the mean weight of babies in a mixed population. The cure rate of cases of typhoid or diphtheria treated by general practitioners cannot be compared with that of hospital cases. Patients treated by private practitioners are mostly those who are in the early stage of disease, conscientious, educated, co-operating and from higher socio-economic strata as compared with those treated in government hospitals who come in advanced stages of disease and belong to lower socio-economic strata. Both types if chosen for generalisation, are biased or selected samples. Prevalence rate of tuberculosis in well-to-do localities or in slums may not give the correct picture of tuberculosis prevalence in Nigeria.

Selection or bias should be avoided in any scientific study when we compare like with the like. The age, sex, social status, stage of disease should be same in both the samples but sometimes it may be deliberate, e.g. we specially want to know the spleen rate in secondary school children only, or the mean blood pressure of people above 50 facing high responsibility or the pulse rate of infants and so on. In such cases, generalisation is to be made about a particular group to compare with another group which is different such as coronary disease in drivers and conductors. Here, the primary objective is to compare like with unlike.

Selection or bias may creep in unconsciously and go unnoticed, hence, it has to carefully be avoided while conducting an experiment, e.g. in enquiries by questionnaire, selection or bias must be suspected. Too keen and intelligent persons might volunteer information and even exaggerate, while those not keen and some with particular habits such as smoking or with disease as gonorrhea, might hide the facts or not reply at all. Unconsciously, sometimes we compare the result of treatment in one group of tuberculosis cases with that of the other, disregarding factors like one or two lungs affected, extent of the disease, age, sex, class, occupation, etc.

In some studies, the effort is made to compare like with the like, but because of unavoidable bias or selection bias, due to number of disturbing factors, one sample may not be similar to the other. Thus, the results may be affected and wrong conclusions drawn, e.g. if the development of babies on breast feeding is compared with those on bottle feeding, it is difficult to get similar samples. Many mothers start supplements in various forms and disturb similarity among the groups.

It is often difficult to rule out bias in retrospective studies and in studies where subjective observations (information given by subjects in the sample) are made. It is easy to avoid bias in prospective studies and in experiment where objective observations (those made by investigator) are made. Subjective bias is ruled out by single or double blind trials.

Wherever possible, a control is a must. Number and type of subjects included in the experiment and technique used for selection should be same for control and experimental groups and decided before choosing the subjects in either group.

Following sampling techniques are employed to choose an unbiased sample. If these techniques are employed, the chance error can be calculated and the probability or relative frequency of getting different results from sample to sample or from sample to those of population can be determined. Hence, samples, thus selected are called *probability samples.* Any unit or member of such samples has a definite probability or chance of being included, e.g. probability of drawing spade ace or any specific card in a pack of playing cards is one in 52 in one draw and of any ace, the probability is 13 in one draw.

Samples can be drawn from the entire population through various procedures. They are as shown below;

i) Simple random sampling
ii) Systematic sampling
iii) Stratified random sampling
iv) Cluster sampling
v) Multistage sampling
vi) Multiphase sampling and
vii) Purposive or quota sampling procedure

3.4 Sampling Procedures

3.4.1 Simple Random Sampling (SRS)

This method is well applicable when the population is small, homogeneous, and readily available, such as patients coming to a hospital or lying in the wards. It is used in experimental medicine or clinical trials like testing the efficacy of a particular drug. The principle here is that every unit of the population has an equal chance of being selected. Hence, this method is also sometimes called unrestricted random sampling. In these sampling procedures, every individual of the population has an equal chance to be selected.

The sample may be drawn unit by unit, either by numbering the units such as persons, families or households of a particular population from the published tables of random numbers. To ensure randomness of selection, one may adopt either lottery or refer to a table of random numbers.

i. Lottery method:
   Suppose 20 patients are needed for an investigation from 50 patients attending the hospital. The procedure is simple and easier for use. All the 50 patients can be given numbers serially from 1, 2 to 50 on 50 pieces of papers of equal size. They can be folded and shuffled. Draw out one and note the number. Replace the piece of paper drawn, reshuffle and draw the second one. Repeat the process till 20 numbers are drawn. Reject the cards that are drawn for second time. The 20patients drawn thus will indicate the patients to be selected for the study. The 20patients selected in this manner constitute the random sample. Similar procedure can be followed for selecting one more group to serve as controls if need is there. Since the procedure is simple and easy for execution, the procedure is known as simple random sampling procedure.

ii. Random number procedure:

The well utilised procedure for the selection of a random sample of subjects is through the use of the *tables of random numbers.* For all the patients of 50, serial numbers can be given (from the table mentioned above) starting from any one of the subjects. This reduces the bias. Total number of subjects is50 which is a two-digit number. Anywhere in the random number tables, two digit numbers can be chosen blindly.

Numbers, less than 50 may be chosen as they are, while those higher than 50 may be divided by 50 and the remainders may be noted as the numbers chosen for the sample. Number higher than 50 could be rejected too, making use of the rows or columns of the random tables thus 20 numbers for sample are chosen.

Example: Select a sample of 10 from a population of 300 female patients attending the National Hospital Abuja: three hundred is the three-digit figure. First three rows of the random table are chosen. The numbers are:

034, 977, 167, 125, 555, 162, 844, 630, 332, 576

The number selected for the sample will be 034, i.e., 34th patient of the subjects and (977/300, remainder 244), 30(630/300, remainder 30), 32(332/300, remainder 32) and 276 (576/300, remainder 276).

Thus, the number of patients to be chosen for the sample is 34, 77, 167, 125, 255, 162, 244, 30, 32 and 276.

If there are some numbers repeated, they can be rejected.

The sample of desired size can be drawn in this way from the population of any size. If the population size is 55,000 and we have to select a sample of 100, start with any five digits of random number table row-wise or column-wise. Every time a number lower than the population has to be chosen. If any higher number is obtained, it has to be divided by the population size, taking remainder as the number selected. If the random number is 75457, the, remainder (75457/55000, remainder is 20457) 20457, has to be taken as the number selected. Continue this process till we get the sample of 100 subjects.

3.4.2 Systematic Sampling

This is a simple procedure and utilised when a complete list of population from which a sample is to be drawn is available. It is more often applied to field studies when the population is large, scattered and heterogynous. Systematic procedure is followed to choose a sample by taking every *n*th house or patient where *n* refers to the sample interval which is calculated by the following formula:

n= (Total population/ Sample size desired)

If 5 per cent of the population is taken as a sample from a population of 5000 households, then *n* is 5000/5 per cent of 5000

$$=\frac{5000}{5\%\ of\ 5000}=\frac{5000}{(0.05)(5000)}=\frac{5000}{250}=20$$

One random number in between 1 to 20 is chosen from random number tables. Since, 20 is the two-digit number, row-wise the first number is ‘03’. In place of random numbers, workers prefer to write serially the numbers 1, 2………, 20 on 20 pieces of papers of equal size. One at random can be chosen which may be 3 or any one of the numbers. If the first random number is 03; then the sample will consist of units with sample numbers 3, 3+20=23, 23+20=43; 43+20=63; 63+20=83; 83+20=103 and so on. Every 20^{th} household after the third must be contacted for the investigation. This type of sampling is flexible and good for use in big cities like Lagos, Abuja, Kano etc.

If one per cent of the sample is to be taken, then

$$n=\frac{5000}{1\%\ of\ 5000}=\left(\frac{5000}{50}\right)=100.$$

We have to contact every 100^{th} household starting with the random number such as five. The subsequent numbers will be 5, 100+5=105, 105+100=205; 205+100=305 and so on.

One in every two households may be taken for study.

$$\text{If } n=2=\frac{5000}{\frac{1}{2}\ of\ 5000}=\frac{5000}{2500}=2.$$

Exercise: select a systematic sample of 10 patients from the available 50 patients for a study for the assessment of the effectiveness of a drug.

Sample number is x= 10

Population size = *X = 50*

sample of,*x* = 10 is equal to 20 per cent of the population.

$$\left(\frac{x}{X}\times 100=\frac{10}{50}\times 100=20\%\right)$$

$$\text{So } n=\frac{\text{x}}{20\%\ of\ X}=\frac{50}{20\%\ of\ 50}=\frac{50}{0.2\ of\ 50}=\frac{50}{(0.2)(50)}=\frac{50}{10}=5$$

Random number chosen (say) from random table is four. Patients to be chosen are with numbers 4, 4+5=9,

9+5= 14; 14+5=19; 19+5=24; 24+5=29;

29+5=34; 34+5=39; 39+5=44; 44+5=49;

The serial number of patients selected with the use of systematic sample is as below:

4, 9 ,14, 19, 24, 29, 34, 39, 44 and 49

Merits of the systemic sampling procedure are:

i) Procedure is simple and convenient for use
ii) Relatively, time to be devoted and labour needed are small: and
iii) If the population is sufficiently large and homogeneous and if the numbering of the subjects is available, this method can provide good results.

An element of randomness is introduced into this kind of sampling by randomly selecting from the first 'n' units, the unit with which it starts. This is referred to as 'Random start', and a sample so chosen is sometimes called "Every 'n' *th* systematic sample".

3.4.3 Stratified Random Sampling

Stratified random sampling procedure is followed when the population is not homogeneous. The population under study is first divided into homogeneous groups called strata and the sample is drawn from each stratum at random in proportion to its size. This procedure gives more representative sample than simple random sampling in a given large population.

Example, the entire population may be divided into four strata based on socio medical groups. From each stratum or socio medical group, sample size fixed may be chosen in proportion to the population of each stratum. If the percentage of the population group-wise or stratum-wise is 20 percent, 40 percent, 30 percent and 10 percent, the sample size can be taken in the same proportion. If a sample size of 50 is to be taken, the distribution of the sample size to be selected stratum wise is 10, 20, 15, and 5 respectively from the 1st to the 4th strata. See details in the following table;

Population by Socio Medical Status					
Details	I	II	III	IV	TOTAL
Percentage of population	20.0	40.0	30.0	10.0	100.0
Distribution of sample size to be selected	10	20	15	5	50

In place of socio medical group as strata, decomposition of population into strata based on age, sex, defined areas, diseases, can also be made. Stratified random sampling gives greater accuracy and better representation to each stratum, compared to simple random sampling.

Another example; Select a stratified random sample of 20 patients from the 200 given below by stratification of diseases
Stratified Random Sample by disease

Strata based on disease					
Details	I	II	III	IV	TOTAL
Number of patients Percentage	100 50.0	60 30.0	20 10.0	20 10.0	200 100.0
Sample size	10	6	2	2	20

Out of 20 patients, number to be selected strata wise are;
Stratum 1 (100/200) x 20 = 50% of 20 =10
Stratum 2 (60/200) x20 = 30% 0f 20 = 6
Stratum iii (20/200) x20 = 10% 0f20 = 2
Stratum iv (20/200) x20 = 10% of 20 = 2

3.4.4 Cluster Sampling

A cluster is a group consisting of units such as villages, wards, workshops or children of a school, factories, slums of a town, etc. sampling procedure is adopted for selection of clusters. Usually, simple random sampling procedure is utilised in the section of clusters. After the selection of clusters randomly, the complete enumeration of the subjects in the cluster is carried out.

Though cluster sampling is simpler and time saving, it is costlier and provides figures with higher standard errors than the other procedures.

Generally, the clusters consist of natural groupings, and if they are geographic regions, the sampling is referred to as area sampling.

3.4.5 Multistage Sampling

This method refers to sampling procedures carried out in several stages using random sampling techniques. This procedure is employed in large scale country wise or region wise surveys. Stages for sampling at the national level are;

	Country	States	Districts	Wards	Villages	Households Or individuals
Stages	1	2	3	4	5	6

Stage wise sampling procedures are to be utilized for the selection of the households or subjects.

3.5.6 Multiphase Sampling

In this method, part of the information is collected from the whole sample and part from the sub sample. In health examination survey among school age or children, all children in the selected school will be surveyed in the first phase. In the second phase for blood examination, or for X-ray of the bones for maturity, a sub sample will be chosen in the second phase or the third phase. Number in the sub-samples in the 2^{nd} and 3^{rd} phase will become successively smaller and smaller. Survey by such procedure is likely to be less costly, less labourious and more purposeful.

3.4.7 Purposive Sampling and Quota Sampling

If some characteristics of the population are known as a result of previous surveys, samples are chosen by purposive selection. As a result, certain features of sample selected purposively are likely to tally with those of the population. Also due to scarcity of time, limitation of the investigation and scarcity of funds, samples of subjects are chosen arbitrarily without random choice. As a result, representative characteristics of the sample are subjective. Inference based on the results of this purposive sample may be limited in nature.

Example of the purposive sampling is QUOTA sampling which is utilised in USA.

As per this method, the investigator or the interviewer is given instruction about certain characteristics (such as age, sex, and socio medical status of the subjects to be selected); the proportion in various groups being chosen to agree with the corresponding proportions in the population. It is possible with this method to have serious discrepancies between the sample and the population. It may arise in respect of some characteristics which have not been taken into account. With this sampling procedure there is nothing to give general confidence about the representativeness of the sample. Many investigator biases are likely to occur.

4.0 CONCLUSION

It is not possible to include each member (sampling unit) of the population in an experimental study or enquiry or examine all the millions people in a particular nation such as Nigeria to find the prevalence of a given disease or the efficacy of a drug in all the patient suffering from such disease. Covering the entire population may be less accurate because a large of number investigators is required to complete such a huge task. As their uniformity and correctness may vary, collection will be costly, time consuming and labourious, hence the necessity for an appropriate sampling technique. The population or universe must clearly be defined before drawing sample. In this unit we shall focus on two of the principal functions of sampling which is the *estimation of population parameters* (mean, proportion etc.) from the sample statistics and precision vis- avis the different sampling procedures.

5.0 SUMMARY

In this unit you have learnt the following:

- Generalised sampling characteristics in research planning and implementation.
- Meaning of precision (sample size and its determination).
- Unbiased characteristics during sampling.
- Different Sampling procedures:
    - Simple Random Sampling
    - Systematic Sampling
    - Stratified Sampling
    - Clustered Sampling
    - Multistage Sampling
    - Multiphase Sampling
    - Purposive Sampling and Quota.

6.0 TUTOR-MARKED ASSIGNMENT

1. Why is sampling important in research?
2. What are the methods available for estimating sample size in
    a) Statistical research problem involving quantitative data?
    b) Statistical research problem involving qualitative data?
3. Write concisely on the following sampling procedures:
    i) Simple Random Sampling
    ii) Systematic Sampling
    iii) Stratified Sampling

iv) Clustered Sampling
v) Multistage Sampling
vi) Multiphase Sampling
vii) Purposive Sampling and Quota.

7.0 REFERENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7thed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

MODULE 2

UNIT 1 DATA COLLECTION

CONTENTS

1.0 INTRODUCTION

In this unit we shall consider the definitions of data, types of data, such as primary and secondary data, meaning of the terms observation and observational units, experiments and surveys, variables, etc.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- list the types of data
- define observational unit
- explain demographic data collection.

3.0 MAIN CONTENT

3.1 What is Data?

A set of values recorded on one or more observational units. For this reason it becomes necessary to appreciate and fully understand the definitions of the terms observation and observational units.

3.2 Observation

This simply refers to an event and its measurements such as Human Heart beat (event) and 72 beats / min (measurement).

3.3 Observational unit

The source that gives observations, such as object, person, etc. is refered to as observational unit. In medical statistics the term individuals or subjects is used more often. Sequel to an appropriately defined research problem and design is data collection. Normally there are two types of data viz;-

i) Primary Data
ii) Secondary data

Primary Data: These are virgin, brand new or fresh data collected for the very first time.

Secondary data: These are data already collected, analysed and processed by someone else. The researcher based on his/her earlier statement of problem and study design has the choice of primary data collection or secondary data collection/compilation.

3.4 Experiments and Surveys

Normally primary data are collected when experiments are conducted at the course of research. During a particular investigation, a variable under test is isolated and its effect or effects is/are measured voluntarily by the researcher. This is what we refer to as experiment.

3.5 Variable

This is a characteristic that takes on different values in different persons, places or things. A quantity that varies within limits such as height, weight, blood pressure, age, etc. it is denoted as X and notation for orderly series as X1, X_2, X_3, X_4, X_5, ……X_n. The suffix n is symbol for number in the series.

3.6 Survey

Is a research method that involves securing information concerning a phenomenon under investigation or study from all or selected number of respondents of the universe concerned? There are several differences between survey and experiments. These can be summarised as shown in the table below;

Table 1.1 Differences between Survey and Experiments

SURVEY	EXPERIMENT
Conducted in descriptive study	Part of experimental research study
Involves larger samples, usually in cross-sectional studies	Involves small samples
Concerned with describing, recording, analysing, and interpreting conditions that exist or existed.	The researcher measures the effects of experiments which he voluntarily conducts.
Mostly applicable to social and behavioral sciences	Mostly applicable to physical and natural sciences.
Suitable for field research	Suitable for laboratory research
Concerned with	

3.7 Collection of Demographic Data

The different sources of collection of demographic data are;

1. Population census
2. Records of vital statistics
3. Records of health departments
4. Records of health institutions
5. Records of health surveys
6. Periodic publication by;

    i. World health organisation
    ii. Registrar general of Nigeria
    iii. Directorate general health services {DGHS}, Abuja F.C.T.
    iv. State health directorates

7. Miscellaneous including other health services agencies and medical establishments like hospitals and nursing homes.

4.0 CONCLUSION

Data is a set of values recorded on one or more observational units. Observation simply refers to an event and its measurements as observational unit which is the source that gives observations. There are two types of data viz; - Primary data and Secondary data. In experiments and surveys primary data are collected when experiments are conducted at the course of research. While Variable takes characteristic that takes on different values in different persons, places or things**.**

5.0 SUMMARY

In this unit we have learnt about the following:

- Types of Data viz;- Primary and Secondary
- Observation
- Observational Unit
- Experiments
- Surveys
- Variables
- Demographic data collection

6.0 TUTOR-MARKED ASSIGNMENT

1 Differentiate between primary and secondary data.
2 What is the meaning of Data?
3 In tabular form only, show the differences between surveys and experiments.
4 How would you collect demographic data?

7.0 REFERENCES/FURTHER READING.

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim&Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6thed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed). Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, AManual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

UNIT 2 POPULATION CENSUS

CONTENTS

1.0 INTRODUCTION

It is an important source of health information. The total process of collecting, compiling and publishing demographic, economic and social data at a specified time or times pertaining to all persons in a country or delimited territory is called a population census (UN handbook of population census methods). In good old days, census was carried out to count people for taxation and army recruitment. Now it is carried out to assess the national needs and plans programmes for the people's welfare. Characteristics features of the census are:

1. Full account of population which include each and every individual and it is carried out at regular intervals. In Nigeria, a census was concluded in the year 1963 and the next in 1973. Since then it is being done every ten years and the last one was conducted in 1983.

2. It pertains to a particular territory and the information is collected by making house to house visit on the specific dates in the first quarter of the first year of each decade.

The census in Nigeria is conducted under which Government is responsible for appointing a census commissioner. In practice, the registrar general of Nigeria is assigned an *ex-officio* charge of census commissioner. The overall census plan is prepared in conformity with ‘principles and recommendation for national population census of the united nation’.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- discuss events that characterise census preparatory phase
- list types of topics covered during census
- explain the uses of population census in public health
- keep records of vital statistics
- explain sample registration system (SRS)
- discuss geometric progression method also as another technique or GP technique.

3.0 MAIN CONTENT

3.1 Census Preparatory Phase

In the preparatory phase, the census commissioner holds elaborate data base conferences and gives a wide publicity to educate people on the importance of census. Preparation of census schedules and their pretesting pilot study and training of enumerators etc. are carried out.

Actual census with a prefixed data (usually march 31) is preceded by house listing operation in which household schedule is canvassed. These operations take approximately one year.

3.2 Types of Topics Covered

Topic covered are essential the following;

i. Total population at the time of census
ii. Age, sex composition and marital status
iii. Language spoken education and economic status
iv. Fertility i.e. number of children born alive between two censuses to all women up to the date of census
v. Citizenship, places of birth urban and rural populations.

The items in (i) and (ii) are collected invariably in any census while those in (iii), (iv) and (v) are often collected. The changes in the schedules are made from census to census depending on the information needed.

3.3 Uses of a Population Census in Health Matter

i. The census determines population of an area which forms basis for calculation of health indices such as birth, fertility, death, morbidity, marriage rates, etc.
ii. All health services are planned and community measures are adopted as per nature and size of the population.
iii. Average rate of growth of population per year is computed from the population enumerated in two census years.
iv. Knowledge of population distribution helps in planning of the other welfare services such as provision of schools, orphanages, food supplies etc.

3.4 Records of Vital Statistics

Another important source of vital statistics in any country is the civil registration system of that country. Civil registration system in almost all countries of the worlds has is very long history. Usually it has a legal backing with varying machineries, systems of reporting, proforma for reporting mode of collecting information etc.

John Graunt (1620-74). The father of vital statistics analysed London bills of mortality i.e. records of baptism, burials and marriages at the churches, kept by the parish authorities and brought out important facts about births and deaths. Williams Farr is another great name who started as a compiler in England in 1839, and made notable contribution during 40 years of his services in development of vital statistics as regards notification, registration, analysis and interpretation.

The history of registration in Nigeria dates back to the 19th century. Initially only deaths began to be registered with the implicit purpose of assessing the health position. Local health officer collected these data and passed them on to the sanitary commissioner of the government of Nigeria. The year 1873 can be regarded as the landmark year of the history of civil registration in Nigeria because it was in that year that Bengal deaths and deaths registration act was passed. This was later adopted by the neighbouring states of Bihar and Orissa, and was used for the formulation of act by the presidencies and provinces, in subsequent years. The registration practices and procedures obtaining in different parts of the country varied widely. Nevertheless, with passage of time vital statistics became the responsibilities of the state health directorates. At the grass root level, the local registrars were drawn from either the police department or the revenue department or the education department. Later on, after the advent of the government (1960-62), registration work was also assigned to the secretaries of village Panchayats in a few states. In urban areas, the responsibility for registering vital events always rested with municipal Corporation or municipalities.

As will be seen from above, there was a great diversity in procedures and practices (including the registration machinery) in different areas of the country. In order to obviate the problems arising from multiplicity of act and rules governing civil registration in different parts of the country, a central legislation of the subject 'registration of births and deaths act' was brought forward in 1969, which makes the registration of births and deaths compulsory within the period specified under rules framed there under by the state governments. It came into forces from April 1, 1970. It may be clarified that the statutory and executive responsibility for registration activities develops under this act with the state authorities. The motivation for bringing forward the central legislation was to introduce standardisation in the concept and definition used in the collection of information. The registrar general of Nigeria, a statutory authority appointed under this act, has the responsibility for issuing general directions, and for co-ordinating and unifying the activities of state chief registrars however, the data collection machinery continued to be the same as before i.e. different state has vested the responsibilities for registration to different departments, e.g., health, revenue, economic, statistics and with the proviso of a minimum of three-tier hierarchy comprising of chief registrar, district registrar and local registrars.

Data on vital events collected from the periphery and the local registrar are passed on to district registrar who, in turn, sends them to chief registrar. The chief registrar of any state or UT has the statutory responsibility for preparing an annual report on the functioning of the civil registration system an annual report on the functioning on the civil registration system, an annual statistical report and for submission of data to the registrar general, statistical Nigeria.

The forms and the period for registration of births, deaths, and still births or the certification of deaths by the attending doctors have been prescribed under rules of the state, which are framed to carry out the purpose of the act. The registration is made compulsory, generally *within 14 days ofbirth* and seven *days of deaths*. A default in reporting can attract fine up to Rs.50.

3.5 Civil Registration System in Nigeria

As in most developing countries has not been functioning satisfactorily and has serious under recording mainly because of illiteracy and lack of much need for having the extract of records. The civil registration system in many states does not net of exceeds even three per cent of vital events though in some of the progressive states like kerala, registration efficiency exceeds 80per cent. It is observed that data obtained from records maintained in routine registers falls short of the actual. In view of the lack of dependable data on civil registration system, and non-receipts of returns from the birth and deaths registrar, the general registrar, Nigeria, has introduced two schemes to monitor the level of registration and its reliability. These are:

i. Sample registration scheme(SRS)
ii. Model registration scheme(MRS renamed as survey of causes of deaths.

3.6 Sample Registration System(SRS)

The main objectives of SRS, introduced in the year 1964, are to provide reliable estimates of births and deaths at the state and national levels of rural and urban areas separately. However, SRS also provides information on other various measures of fertility and mortality. The field investigation of SRS consists of different enumeration of births and deaths in sampled villages/urban blocks by a resident part-time enumerator preferably a teacher and independent retrospective surveys by a full time supervisor.

The data thus obtained by these two sources are unmatched and partially matched events are re-verified in the field to get an unduplicated and complete count of all the events. SRS now covers the entire country.

The sample design is random stratified sampling in the rural areas. Stratification is done on the basis of natural division and population size classes. Each natural division within a state has been considered as a stratum and further stratification has been done by grouping the villages into population size classes. In the urban areas, stratification has been done on the basis of population size of cities and towns. The sample unit in rural areas is either a village (if the population is less than 2000) or a segment of a village (if the population is 2000 or above). The sample unit in urban areas is a census enumeration block with an average population size of 750.SRS coves approximately six million populations in 4149 rural and 1873 urban i.e. a total of 6022 sample units.

An enumerator is required to send a monthly report to the state headquarters in the state headquarters in the first week of the following month. On the basis of the monthly reports received from the sample units, the state headquarters are required to prepare a consolidated monthly report and forward the same to the office of the registrar general by the end of the following month.

At the national level, the vital statistics division of the office of the registrar general, Nigeria, co-ordinates the implementation work, formulates and prescribes necessary instructions and guidance undertakes tabulation and analysis of data and their dissemination. Latest birth, death and infant mortality rates are given

Ref: sample registration system 1984, vital statistics division, office of the registrar general, Nigeria, ministry of home affairs, Abuja

3.7 Natural Increase Method

Natural increase implies the difference between increase population due to births and immigration and decrease due to deaths and emigration over a period of time- for instance, one year or six years. This can be added to the previous census population to get the mid-term population of a particular year up to which the net increase has been calculated.

This requires reliable recording as done in developed countries like England and Wales but not in developing countries like India. Moreover, by this method, the post censal or future population cannot be projected.

3.7.1 Arithmetical Progression or AP method

Here it is assumed that there is equal increase each year and average increase per year is calculated in the inter-censal period of 10 years such as between 1981 and1991.

Population of any year or up to a particular month=last census population + period in years and month after the last census month and year*yearly increase.

This method is adopted for its being simple to find inter-censal population. It is like simple interest and pays no regard to fluctuation in births, deaths, immigration and emigration. To calculate the health indices for any year mid-year population is used as denominator, though population for any month in the year can be calculated if need be. The formula applied is:

$P_{t=}P_0$+rt
Where t is the period in years after the last census, P_t is the population at the required time i.e., t years after the last census, P_0 is the population of last census and r is the annual increase rate.

Example
The population of Abuja as per march 1, 1981 was 62,200,000 and as per march 1st,1991. Census was 94,21,000. Calculate the mid-year population in 1996 and 2001.

Increase in 10 years =94,21,000-62,20,000=32,01,000
Average increase per year =32,20,100
To determine the mid-year population of 1996, add the increase at the rate of 3,20,100 for five years and four months from 1st march, 1991 to 30th June,1996.

A. Mid-year population in 1996

=94,21,000+3,20,100*5+3,20,100*1/3(1/3 means 4 months from 1st to 30th June) =1,11,28,200

Similarly, to calculate mid-year population in year 2001, add the increase at the rate of 3,20,100 for 10 years and four months from 1st march, 1991 to June 1991.

B. Mid- year population in year 2001

= 94,21,000+3,20,000*10+3,20,100/3
=94,21,000+3,200,000+1,06,700
=1,27,28,700

3.7.2 Geometrical Progression or GP Method

It is based on the principles, population begets population. In AP method constant increase throughout was assumed while in this, assumption is made that the percentages or per person rate of growth remains constant. It is calculated like compound interest. If a population grows from one million to three millions in one year, next year it grows from three million to nine millions and so on i.e., as per geometric law of growth.

Suppose p_0 is the population of any census year and r is the annual increase per person in inter-censal years and p_t(population after t years) is to be found. The formula applied is:

$P_t=P_o(1+r)^t$

At the end of the year it becomes $P_0(1+r)^t$
At the end of two years=$P_O(1+r)^t$

At the end of ten years=$P_0(1+r)^t$
So, at the end of t years=$P_0(1+r)^t$
It is easy to calculate with the help of log tables.

Examples

i. Census population of Abuja on 1st march, 1981 and 1991 was 62,20,000 and 94,21,000, respectively. Calculate the mid-year population of Abuja in 1996 and 2001.

If 'r' denotes the rate at which the population increases in geometric progression per year. The first step would be to calculate 'r'. Population has gone up in 10 years from 62,20,000 to 94,21,000.

Population on 1st march, 1991= population on 1st march, 1981*$(1+r)^{10}$
94,21,000=62,20,000*$(1+r)^{10}$

A. taking logarithm

log94,21,000= log 62,20,000+10log(1+r)
log9.421+3log 10= log622+4log10+10log(1+r)
3.9741+3=2.7938+4+10 log (1+r)
Log (1+r) =0.01803
After march 1991 in five years four months, the mid-year population in 1996=94,21,000 $(1+r)^{5+1/3}$
Taking log, we get log (mid-year population in1996)
=log 94,21,000+ {5+1/3} log(1+r)
= 6.9741+16/3*0.1803
=7.0702
Taking antilog, mid-year population becomes
= 1,17,60,000

B. similarly, the mid-year estimated in 2001

=94,21,000 $(1+r)^{10+1/3}$
Taking logarithms
Log (mid-year population in 2001) = log 94,21,000+ {10+1/3} log (1+r)
=6.9741 +31/3* 0.01803
=7.1604
Taking antilog, we get mid-year population in 2001
= 1,44,60,000

ii.

a. the population of Nigeria as per 1st march, 1981 was 685 million and as per 1st march, 1991 census, population was 844 million. Calculate the mid-year population in 1996 and 2001

Increase in 10 years = 844-685= 159 million

Average increase per year = $\frac{159}{10}$ =15.9million

To determine the mid-year population of 1996, add the increase at the rate of 15.9 million for five years from 1st march, 1991 to 1st march, 1996.

Mid-year population in 1996= 844+15.9*5+15.9/3=923.5 million +5.3

=928.8 million

Similarly, to calculate the mid-year population in 2001 add 15.9 per year from 1st march, 1991 to 1st march, 2001plus increase of 4 months till mid-year at the end of June.

Mid-year population in 2001

= 844+15.9*10+15.9/12*4=948.3 million

b. Census population of Nigeria on 1st march, 1981 and 1st march, 1991 was 685 million and 844 million respectively. Estimate the mid- year of Nigeria for 1996 and 2001.

If 'r' denotes the rate at which the population increases at geometric progression per year, the first step would be to calculate for 'r'. Population has gone up to 10years from 685 million to 844 million

Population on 1st march, 1991= population on 1st march 1981*$(1+r)^{10}$

84,40,00,000=68,50,00,000 $(1+r)^{10}$

Taking logarithm on both sides, we get

Log 84,40,00,000=log 68,50,00,000+10 log (1+r)

= log (844*1000000) =log (685*10,00,000) + 10 log (1+r)

= log844+6 log 10= log 685+ 6 log 10+ 10 log (1+r)

2.9263=2.8357+ 10 log (1+r)

10 log (1+r) = 0.0906

Log (1+r) = 0.00906

After March 1991 in five years the mid-year population in 1996=84,40,00,000$(1+r)^5$

Taking log,

Log (mid-year population in 1996)

=log 844+ 6 log 10+ 5log (1+r)

= 2.9263 + 6 + 5*0.00906

= 8.9716

Taking antilog, the mid-year population of Nigeria in 1996 comes to 936.7 million.
Similarly, the mid-year estimated population of 2001

$$= 84,40,00,000\ (1+r)^{10}$$

Taking log
Log(mid-year population in 2001)

$$= \log 844 + 6\log 10 + 10 \log (1+r)$$
$$= 2.9263+6+10*0.00906$$
$$= 9.0169$$

Taking antilog, the mid-year population of Nigeria in 2001 comes to 1,04,00,00,000 i.e., 1040 million.

The reader is notified that the above population calculated is till March 1996 and 2001. He should add population increase for four months by the same method using t=1/3 years and add to the above calculated value to get exact mid-year population till the end of June 1996 and 2001, respectively.
Based on principle of compound interest, geometrical progression or increase is a better method to follow. It gives higher estimate of post-censal and lower estimate of inter-censal period than AP method. The same is noticed in example 1 and 2, where post-censal population of Abuja or of Nigeria, respectively will be found higher, and in the inter-censal, will be found lower if calculated in 1976.

3.7.3 Growth of Population (Natural Increase)

It is often required to calculate the yearly growth rate of population in family planning and demography. It may be found by subtracting annual death rates from birth rate per thousand. Sometimes, it is expressed in percentage.

The population in Nigeria has been rising by 2.23% per year in 1981-91 decenniums against the annual rise of 2.22% and 1.89% in the previous two decades i.e.1971-81 and 1961-71 respectively.

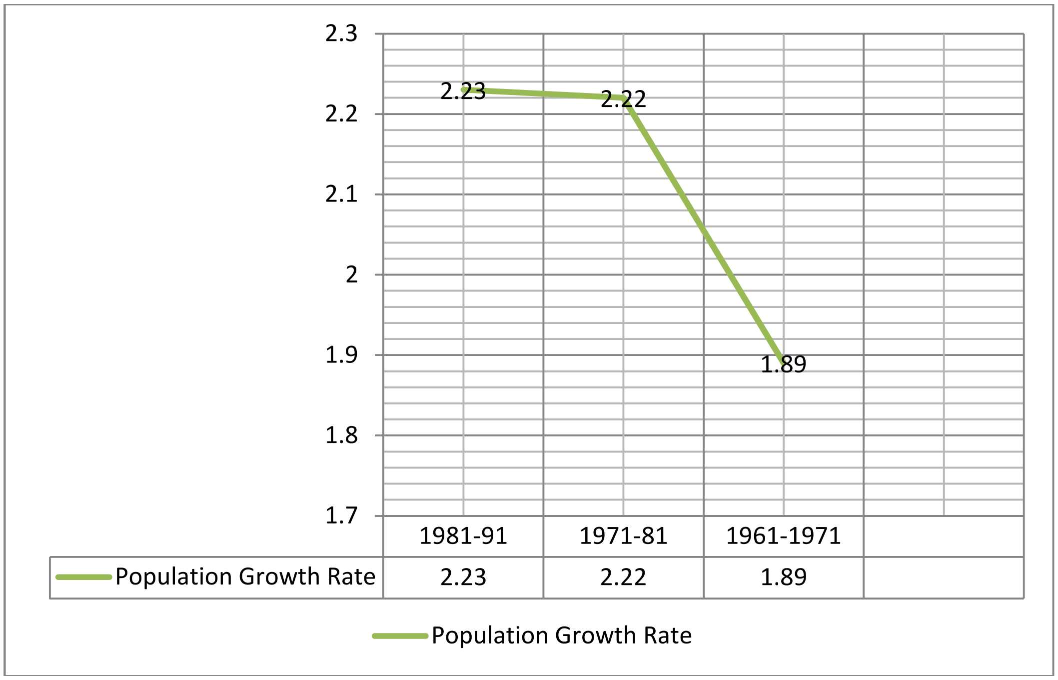


Fig.2.1: Population Growth Rate

Table 2.1: Growth Rate

YEAR	PERSONS	MALE	FEMALE	SEX (MALE/FEMALE)
1981	685.18	357.38	333.77	1071
1986	761.07	392.79	368.28	1066
1991	837.28	431.33	405.59	1062

It is seen from the table and figures that declines death rate is much smaller as compared with increases in birth rates; nearly 16.06 million people were being added every year in the decennium 1981-91. In 1996, estimated population was 500 million. With birth rate of 40 there were 20 million births and with death rates reduced to 16, there were 8 million deaths. Thus, the increase in population went up by 12 million giving growth rate of 24 per thousand. As per sample registration scheme in Nigeria, in 1993, the crude births rate was 28.5, crude death rate 9.2 and natural growth rate was 19.3 per thousand.

3.7.4 Population Density

It is defined as number of people living in one square kilometres area. It is a good indication to assess the slow or rapid increase in congestion due to population growth over different period of time. With an unplanned increase, the health of people is affected, thereby, morbidity and mortality rates goes up.

Population density has to be compared among different towns or regions to study the congestion.

Example

Area of Nigeria is 31,60,789 sq. kilometres and the population in 1981 and 1991 censuses was 683 and 843 million, respectively. Calculate the rise in population density.

Density in 1981= $\frac{68,33,29,097}{31,60,789}$=216/sq. km

Density in 1991= $\frac{84,39,30,861}{31,60,789}$= 267/sq. km

Rise in density = 267-216=51/sq. km in the decade

Density in 1901, was 77 per sq. km. with increase in population decade wise, it rose to 82,81,90,103,117,142,177 and 216 in 1981. In the year 1991, it went up to 267 (source: provisional population tables. Census of Nigeria)

Density of population in a town may be calculated as person per room to determine overcrowding. Density ratio of over two persons per room is indicative of overcrowding *studies have revealed that general morbidity and infant mortality rates are correlated with overcrowding*s

4.0 CONCLUSION

Census is an important source of health information. The total process of collecting, compiling and publishing demographic, economic and social data at a specified time or times pertaining to all persons in a country or delimited territory is called a population census (UN handbook of population census methods). It entails full account of population which include each and every individual and it is carried out at regular intervals. In Nigeria, a census was concluded in the year 1963 and the next in 1973. Since then it is being done every ten years and the last one was conducted in 1983.it pertains to a particular territory and the information is collected by making house to house visit on the specific dates in the first quarter of the first year of each decade. The overall census plan is prepared in conformity with ‘principles and recommendation for national population census of the united nation’.

5.0 SUMMARY

In this unit we have learnt the following:

- Events that characterise census preparatory phase
- Types of topics covered during census
- Uses of Population census in public health
- Records of vital statistics
- Civil Registration System in Nigeria with respect to Census
- Sample Registration System (SRS)
- Natural increase method as example of appraisal technique used in census
- Arithmetic Progression Method as another technique or AP Technique
- Geometric Progression Method also as another technique or GP Technique
- Population Growth Technique or Natural Increase method
  Population Density Technique

6.0 TUTOR-MARKED ASSIGNMENT

Write exhaustively on census and its Modus Operandi in your Country of origin**.**

7.0 REFERENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum’s outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

UNIT 3 LIFE TABLE

CONTENTS

1.0 INTRODUCTION

Under the measures of mortality descried in the previous chapter, an important measure; '**Expectation of life'** at birth or at any other age was not discussed. This involves the use of life table which is a particular way of expressing death rate experienced by a particular population at a particular period. Life table is a special type of cohort analysis which takes into account the life history of **hypothetical group** or **cohort** of people that decrease gradually by death till all the member of the group died. It is a simple device not only for mortality but for other vital events like natality, reproduction, chances of survival, etc.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- discuss life table
- state how to apply life table
- construct a life table.

3.0 MAIN CONTENT

3.1 Uses and Application of Life Table

1. To find the number of **'survivors'** out of 1000 or 10000 or over, birth or at any age thereafter say:
i. At the age of five to find the number of children likely to enter primary school.

ii. At the age of 15 to find number of women entering fertile period or find the number of adolescents entering the community.
iii. At the age of 18 to find the number of persons who become eligible for voting; at 55 or 58 or 60 who become eligible for pension: and at 65 and over to assess the geriatric problems and solution like income tax benefits, railway concession, etc.
iv. At the age of 45 or above to find the number of women reaching menopause.
v. Survival at different ages of women in fertile period and birth of female children from the age 0 to 15 which helps in finding age wise fertility rate, net reproduction rate, etc.

2. To estimate the number likely to die after joining service till retirement, helping, thereby in budgeting for payment towards risk or pension.

3. To find expectation of life or longevity of life at birth or any other age. Increase in longevity of life means reduction in mortality. Thus, life table is another method applied to compare mortality of two places, periods, professions or groups.

4. To find survival rate after treatment in a chronic disease like tuberculosis and cancer or after cardiac surgery like bypass or heart transplantation by modified life table technique.

In short, life table helps to project population estimates by age and sex. William Farr called it a **bio meter of population**.

Table 3.1 indicates the longevity or expectation of births or deaths by sex of different state of Nigeria worked out from 1986-90

Table 3.1: Longevity or Expectation of Births or Deaths

States	Males	Females	Person
ADAMAWA	58.2	60.4	59.1
ABIA	53.7	54.2	53.6
BENUE	55.7	53.6	54.9
BENIN	57.0	58.8	57.7
BORNO	62.2	62.2	62.2
CROSS RIVER	62.4	62.8	62.8
DELTA	60.4	62.6	61.1
ENUGU	66.8	72.3	69.5
EDO	53.7	53.0	53.0
ZAMFARA	61.2	63.5	62.6

3.2 Construction of a Life Table

To construct a life table, two things are acquired;

1. *Population living at all individual ages in a selected year.*
2. *Number of deaths that occurred in these ages during the selected year.*

Selected year should be the most recent one for which accurate statistics is available.

Basic element of life table is q_x i.e. probability of dying from age x_0 to x_1, x_2, x_3………to x_n. It is computed per person for each year or life span of years. Calculation of death rate is based on mid-year population. To calculate the mid-year population for first 24 month of life, needs special attention as the death do not occur uniformly in this period.

From third year onwards, mid-year population is calculated easily because deaths are more or less evenly distributed. Half the deaths occur by 30th June, so half of death can be added to the population of cohorts that started life (1_x) at any age to get mid-year population. Person-years lived (L_x) is also easy to calculate. It will be equal to $1_x + 1/2d_x$.

During infancy and early childhood, mortality changes rapidly with age even within the interval of a single year.

Hence, presumption of equal distribution of deaths in two halves of first two years is not valid.

To calculate mid-year population following assumptions for two halves of a year are considered useful (x implies age, in the first year of life)
X_{0-6m} : X_{6m-12m} : : 0.7 : 0.3 (in the first year of life)
X_{12m} : X_{18-24m} : : 0.6 : 0.4 (in the second year of life)
$X_{2y-2.5y}$: $X_{2.5y-x3y:}$: 0.5 : 0.5 (in the third year of life onwards)

Computation of mid-year population in any year or span of years as well as of person- years contributed to 1_x by $d_{x,}$ is done by demographers on these assumption depending on the infant toddlers mortality and availability of records, etc.

P_x i.e., probability of survival is equal to 1-q_x because total probability is one (1).

Imagine a cohort of 1,00,000 new-borns, starting life together. Subject them (1_0) and the survivors at each age (1_x) to the mortality rates of the selected year, till all members of the cohort die.

Table 3.2 Mortality Rates

Age x 1	Number Started life 1_x 2	Number Died	No.of person-years lived	Total Person-years lived	Expectation Of life
		D_x 3	L_x 4	T_x 5	E_x 6

0 1 2 3 4 5 6	1,00,000 90,000 87,300 85,554 83,843 82,585 81,594…. And so on	10,000 2700 1746 1711 1258 991	92,500 88650 86427 84699 83214 82090	92500 181150 267577 352276 435490 517580	$\underline{EL_x}$ $\underline{L_x}$

Indicate age x by suffixing the number of years at the foot such as $x_{o,}$ x_1............x_n ,survivors 1_x at different ages as $l_{o,}$ $l_{1,}$ $l_{2,}$ ....$l_{n,}$ number died d_x as $d_{1,}$ $d_{2,}$ ………$d_{n,}$ years lived l_x as l_1,l_2, …….l_n and total years lived as $t_x$as $t_{1,}$ $t_{2,}$ ……….., $t_{n.}$

Column x indicates the age at which cohorts start life such as at birth indicated by 0 or at any age after that such at age one, age two and so on.

Column 1_x gives the number that *started life* at any particular age such as 100000 at birth, 90000survivors at age two and so on.

Column d_x is the number died in each year, such as 10000in the first year from age 0 to age 2700 in the second year and 1746 in the third year and so on. This number is found at the specific mortality rate at each year, e.g., in the first year from age 0 to age one, 10000 died at the infant mortality rate of 100 per 1000. Deaths in the 2^{nd} ,3^{rd} ,4^{th} ,5^{th},6^{th}, year of age are calculated in the above table at the specific deaths rate of 30, 20, 20, 15 and 12 respectively.

L_x column gives the estimated total number of person years *lived* by the cohort at each age. It will always be more than the survivors at the end of any particular year because all do not die in the beginning of the year. They live for some days or months in each year before death, e.g., in the first year, 10000 died and 90000 survived (1,00,000-10000) in this year most deaths occur in the first week and first month and less and less later, even in the countries with high infant mortality rate such as 100.it may be presumed that 10000 before death, lived for three months on an average so they contributed (3/12*10000) 2500 years. Thus, total years lived L_o in the first year will be 92,500 (90000+2500). In subsequent years, it is presumed that a person lives for half a year on an average before death. L_1 or number of years lived up to the end of second year will be 87,300+1/2 of 2700=88650 and so on.

Column T_x gives the total number of years lived till any age, e.g., 1,00,000 that started `life, lived 90,000+2,500=92500 years up to age one as entered under column opposite age one, same as $L_{1.}$

Up to the age two, total number of years lived will be L_1 till end of first year 92500 + $L_{2,}$ i.e., years lived till the end of second year, 8,86,550= 1,81,150 as entered opposite age two.

Up till the end of five years, T_5=L_1 + L_2 +L_3 + L_4 +L_5 (92500 + 88650 + 86427 + 84699 + 83214= 4,35,490). T_{50} or total years lived up to age 50 will be the total years lived by the starters till 50 years passed (L_1 + L_2 + $L_{3,}$...............L_{50}). final T_n, when all cohorts die off, will give the number of years lived by 1,00,000 that started life (l_0).

The mean expectation or longevity of life, i.e., average number of years a person is likely to live at age 0 or at any age after that is denoted by e_x. this is obtained by subtracting the year already lived from the final total number of years lived and dividing the balance by the number of starters at the age, longevity is desired to be calculated.

E_x=<u>EL_x = sum of person-years lived till age x</u>

$l_x$number of starters

mean or average expectation of life (e_x) can also be found at any age (x) from the survivors column l_x by the formula-

e_x= <u>sum of l_x column excluding those starting life + 1/2</u>

number that started life

sum of l_x gives the total person years lived. dividing by the number of starters. Gives mean years lived per person.

Half year is added as an average period lived after completion of the last year.

Longevity of life can also be calculated from the column of died, d_x, by the formula

$E_{x=}$<u>d_{x0}*1/2+d_{x1}*$1^{1/2}$+d_{x2}*$2^{1/2}$+…….+d_{xn}*$n^{1/2}$</u>

Number That Started life

Number of years lived by the deceased in each group are found by multiplying the deaths by age +1/2years i.e., d_x* (x+1/2). The d_{x0} means the number of death that occurred in the first year of life from age 0 to age 1. When multiplied by 1/2, it gives the number of years lived by the deceased in the first year. The d_{x1}, means the number of deaths in the second year. When multiplied by 1 ½ it gives the number of years lived by the deceased I n the second year, and so on. Thus, d_x column can also give the number of years lived by all before death. This is more laborious than the calculation from 1_x.

Life tables can separately be prepared for sexes, professions, places and periods and comparisons made on the basis of **life expectancy**. Such comparisons are independent of age and sex composition.

Cohort analysis by life table can be used as a measure of total fertility. Start with a cohort of women in child bearing age and subject the age groups to different fertility rates of most recent years of which fertility are available. Net reproduction rate (RRR) described is also calculated by life table method.

Population and death for ages 0-4 are given in the table below suppose 10000 children start life, make a life table for them and find the number surviving at each successive birthday till the 5^{th}.

Table 3.4 Life Table

AGE	POPULATION	DEATHS
0-	4151	70
1-	4792	16
2-	4797	6
3-	4998	4
4-5	4798	4

Solution

In the first of life, all the 70 deaths did not occur on the first day. Probably, $7/10^{th}$ died by mid-year, therefore mid-year population of first year would have been 4151 + 7/10*70=4200. Then the chance of dying per person i.e.,

$D_x = \frac{70}{4200} = 0.0167$. Thus, 167 ($d_x$) out of 10000 would actually die,

So the number of survivor entering life for 2^{nd} year is 10000-167= 9833

After the first year, we presume that half the death occurred before and half death occurred after six months. When 16 deaths occurred, they lived eight years more. These eight years has to be added to those that reached second year, i.e., 4792+8=4800 chance of dying per person = 16/4800 = 0.0033. Multiply 9833 by 0.033 to find number died at the second year at this rate, 9833*0.033=32. Population surviving to start the third year of life will be 9833-32=9801.

Calculate the number of deaths out of 9801, and similarly, the number of those starting life in the fourth and fifth years of life and find the number that survive to start in the sixth year and fifth birthday

4.0 CONCLUSION

Expectation of life' at birth or at any other age involves the use of life table which is a particular way of expressing death rate experienced by a particular population at a particular period. Life table is a special type of cohort analysis which takes into account the life history of **hypothetical group** or **cohort** of people that decrease gradually by death till all the member of the group died. It is a simple device not only for mortality but for other vital events like natality, reproduction, chances of survival, etc. In life table we address issues like Uses of Life Table, Applications of Life Table, and Construction of Life Table.

5.0 SUMMARY

In this unit we have learnt about the following

- Uses and application of life table
- Construction of life table

6.0 TUTOR-MARKED ASSIGNMENT

Explain the two things required to construct a life table.

7.0 REFERENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica, Sheringham, Kanwal,Kalim&Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

MODULE 3

UNIT 1 DATA PREPARATION

CONTENTS

1.0 INTRODUCTION

The data, after collection, has to be prepared for analysis. The collection data is raw and it must be converted to the form that is suitable for the required analysis. The results of the analysis are affected a lot by the form of the data. So, proper data preparation is a must to get reliable results.

2.0 OBJECTIVE

By the end of this unit, you will be able to:

- explain the following steps as they concern data preparation:
  questionnaire checking
  editing

coding
classification
tabulation
graphical representation
data cleaning
data adjusting
missing values and outliers
various types of analysis.

3.0 MAIN CONTENT

3.1 Questionnaire Checking

When the data is collected through questionnaires, the first step of data preparation process is to check the questionnaires if they are acceptable or not. This involves the examination of all questionnaires for their completeness and interviewing quality. Usually this step is undertaken at the time of data collection. If questionnaire checking was not done at the time of collection, it should be done later. A questionnaire may not be acceptable if:

1. It is incomplete partially or fully.
2. It is answered by a person who has inadequate knowledge or does not qualify for the participation.
3. It is answered in such a way which gives the impression that the respondent could not understand the questions.
4. If sufficient number of questionnaires is not accepted the researcher may like to collect more data.

3.2 Editing

Editing of data is a process of examining the collected raw data (especially in surveys) to detect errors and omissions and to collect these when possible. As a matter of fact, editing involves a careful scrutiny of the completed questionnaire and/or schedules. Editing is done to assure that the data are accurate, consistent with other fact gathered, uniformly entered, as completed as possible and have been well arranged to facilitate coding and tabulation.

With regard to points or stages at which editing should be done, one can talk of field editing and central editing. *Field editing* consists is the review of the reporting forms by the investigator for completing (translation or rewriting) what the letter has written in abbreviated and/or in illegible form at the time of recording the respondent' responses. This type of editing is necessary in view of the fact that individual writing styles often can be difficult from others to decipher. This sort of editing should be done as soon as possible after the interview, preferably on the very day or on the next day. While doing field editing, the investigator must restrain himself and must not correct error of omission by simply guessing what the informant would have said if the question had been asked.

Central editing should take all forms or schedules have been completed and returned to the office. This type of editing implies that all forms should get a thorough editing by a single editor in a small study and by a team of editors in case of a large inquiry. Editor(s) may correct the obvious errors such as an entry in the wrong place, entry recorded in months when it should have been recorded in weeks, and the like. In case of inappropriate on missing replies, the editor can sometimes determine the proper answer by reviewing the other information in the schedule. At time, the respondent can be contacted for clarification. The editor must strike out the answer if the same is inappropriate and he has no basis for determining the correct answer or the response. In such a case an editor entry of 'no answer' is called for. All the wrong replies, which are quite obvious, must be dropped from the final result, especially in the context of mail surveys.

Editors must keep in view several points while performing their works:

(a) They should be familiar with instructions given to the interviewers and coders as well as the editing instructions supplied to them for the purpose.
(b) While crossing out an original entry for one reason or the other, they should just draw a single line on it so that the same may remain legible.
(c) They must make entries (if any) on the form in some distinctive color and that too in a standardised form.
(d) They should initial all answers which they change or supply.
(e) Editor's initials and the date of editing should be placed on each completed form or schedule.

3.3 Coding

Coding refers to the process of assigning numerals or other symbols to answers so that responses can be put into a limited number of categories or classes. Such classes should be appropriate to the research problem under consideration. They must also possess the characteristic of exhaustiveness (i.e. there must be a class for every data item) and also that of mutual exclusively which means that a specific answer can be placed in one and only one cell in a given category set. Another rule to be observed is that of unit dimensionality by which is meant that every class is defined in terms of only one concept.

Coding is necessary for efficient analysis and through it the several replies may be reduced to a small number of classes which contains the critical information required for analysis. Coding decisions should usually be taken at the designing stage of the questionnaire. This makes it possible to precode the questionnaire choices and which in turn is helpful for computer tabulation as one can straight forward keg punch from the original questionnaire. But in case of hand coding some standard method can be used. One such standard method is to code in the margin with a coloured pencil. The other method can be to transcribe the data from the questionnaire to a coding sheet. Whatever method is adopted; one should see that coding errors are altogether eliminated or reduced to the minimum level.

3.4 Classification

Must research studies result in a large volume of raw data which must be reduced into homogeneous groups if we are to get meaningful relationships. This fact necessitates classification of data which happens to be the process of arranging data in groups or classes on the basis of common characteristics. Data having a common characteristic are placed in one class and in this way the entire data get divided into a number of groups or classes. Classification can be one of the following two types, depending upon the nature of the phenomenon involved:

(a) Classification according to attributes: As stated above, data are classified on the basis of common characteristics which can either be descriptive (such as literacy, sex, honesty, etc.) or numerical (such as weight, height, income, etc.). Descriptive characteristics refer to qualitative phenomenon which cannot be measured quantitatively; only their presence or absence in an individual item can be noticed. Data obtained this way on the basis of certain attributes are known as *statistics of attributes* and their classification is said to be classification according to attributes. Such classification can be simple classification or manifold classification. In simple classification we consider only one attribute and divide the universe into two classes – one class consisting of items possessing the given attribute and the other class consisting of items which do not possess the given attribute. But in manifold classification we consider two or more attributes simultaneously, and divide that data into a number of classes (total number of classes of final order is given by 2*, where n = number of attributes considered) *. Whenever data are classified according to attributes, the researcher must see that the attributes are defined in such a manner that there is least probability of any doubt /ambiguity concerning the said attributes.

(b) Classification according to class interval. Unlike descriptive characteristics, the numerical characteristics refer to quantitative phenomenon which can be measured through some statistical units. Data relating to Heart beat, blood pressure, blood-sugar, age, weight, etc. come under this category. Such data are known as statistics of variables and are classified on the basis of class intervals. For instance, persons whose ages are within 40 - 50 years can form one group; those whose ages fall between 51- 61 years shall form another group. Each group of class –interval, thus, has an upper limit as well as a lower limit which are known as class limits. The difference between two class limits is known as class magnitude. We may have classes with equal class magnitudes or with unequal class magnitudes. The number of items which fall in a given class is known as the frequency of the given class. All the classes or groups, with their respective frequencies taken together and put in the form of a table, are described as group frequency distribution or simply frequency distribution. Classification according to class intervals usually involves the following three main problems:

i. How many classes should be there? What should be their magnitude? There can be no specific answer with regard to the number of classes. The decision about this calls for skill and experience of the researcher. However, the objective should be to display the data in such a way as to make it meaningful for the analyst. Typically, we may have 5 to 15 classes. With regard to the second part of the question, we can say that, to the extent possible, class –intervals should be of equal magnitude, but in some cases unequal magnitudes may result in better classification. Hence, the researcher's objective judgement plays an important part in this connection. Multiples of 2, 5, and 10 are generally preferred while determining class magnitudes. Some statisticians adopt the following formula, suggested by H.A Surges, determining the size of class interval:

$$I = R/ (1+3.3 \log N)$$

Where
I = size of class interval:
R = Range (i.e. difference between the values of the largest item and the smallest item among the given items):
N = Number of items to be grouped.

It should also be kept in mind that in case one or two or few items have very high or very low values, one may use what are known as open ended intervals in the overall frequency distribution. Such intervals may be expressed like under Rs. 500 or Rs. 10001 and over. Such intervals are generally not desirable, but often cannot be avoided. The researches must always remain conscious of this fact while deciding the issue of the total number of class intervals in which the data are not to be classified.

ii. How to choose class limits?
While choosing class limits, the researcher must take into consideration the criterion that the mid-point (generally worked out first by taking the sum of the upper limit and lower limit of a class and then divide this sum by two) of a class- interval and an actual average of items of that class interval should remain as close to each other as possible. Consistent with this, the class limit should be located at multiples of 2, 5, 10, 20, 100 and such other figures. Class limits may generally be stated in any of the following forms:

Exclusive type class intervals: they are usually stated as follows:
10-20
20-30
30-40
40-50

The above intervals should be read as under:
10 and under 20
20 and under 30
30 and under 40
40 and under 50

Thus, under the exclusive type class intervals, the items whose values are equal to the upper limit of a class are grouped in the next higher class. For example, an interval whose value is exactly 30 would be put in 30-40 class intervals and not in 20-30 class intervals. In simple words, we can say that under exclusive type class intervals, the upper limits of a class interval is excluded and items with values less than the upper limit (but not less than the lower limit) are put in the given class interval.

Inclusive type class intervals: they are usually stated as follows:
11-20
21-30
31-40
41-50

In inclusive type class intervals, the upper limit of a class interval is also included in the concerning class interval. Thus, an item whose value is 20 will be put in 11-20 class intervals. The stated upper limit of a class interval 11-20 is 20 but the real limit is 20.99999 and as such 11-20 class interval really means 11 and under 21.

When the phenomenon under consideration happens to be discrete one (i.e., can be measured and stated only in integers), then we should adopt inclusive type classification. But when the phenomenon happens to be a continuous one capable of being measured in fractions as well, we can use exclusive type class intervals.

iii. How to determine the frequency of each class?
This can be done either by tally sheets or mechanical aids. Under the technique of tally sheet, the class groups are written on the sheet of paper (commonly known as the tally sheet) and for each item a stroke (usually a small vertical line) is marked against the class group in which it falls. The general practice is that after every four small vertical lines in a class group, the fifth line for the item falling in the same group is indicated as horizontal line through the said four lines and the resulting flower (~~IIII~~) represents five items. All this facilitates the counting of items in each one of the class groups. An illustrative tally sheet can be shown as under:

Table 1.1: An Illustrative Tally Sheet for Determining the Number of 70 Families in Different Income Group

Income groups (rupees)	Tally mark	Number of families or (classfrequency)
Below 400	~~IIIII IIIII~~ III	13
401-800	~~IIIII IIIII IIIII IIIII~~	20
801-1200	~~IIIII IIIII~~ II	12
1201-1600	~~IIIII IIIII IIIII~~ III	18
1601 and above	~~IIIII~~ II	7
Total		70

Alternatively, class frequencies can be determined, especially in case of large inquires and surveys, by mechanical aids i.e., with the help of machines viz., sorting machines that are available for the purpose. Some machines are hand operated, whereas other work with electricity. There are machines which can sort out cards at a speed of something like 25000 cards per hour. This method is fast but expensive.

3.5 Tabulation

When a mass of data has been assembled, it becomes necessary for the researcher to arrange the same in some kind of concise and logical order. This procedure is referred to as tabulation. Thus, tabulation is the process of summarising raw data and displaying the same in compact form (i.e. in the form of statistical tables) for further analysis. In a broader sense, tabulation is an orderly arrangement of data in columns and rows.

Importance of tabulation

1. It conserves space and reduces explanatory and descriptive statement to a minimum.
2. It facilitates the process of comparison.
3. It facilitates the summation of items and the detection of errors and omissions.
4. It provides a basis for various statistical computations.

Tabulation can be done by hand or by mechanical or electronic devices. The choice depends on the size and type of study, cost considerations, time pressures and the availability of tabulating machines or computers. In large inquiries, we usually use mechanical or computer and in small inquiries where the number of questionnaires is small and they are of relatively short length, Hand tabulation may be done using the direct tally, the list and tally or the card sort and count methods.

Tabulation may be classified as simple and complex tabulation. The former type of tabulation gives information about one or more groups of independent questions, whereas the latter type of tabulation shows the division of data in two or more categories and as such is designed to give information concerning one or more sets of inter-related questions. Simple tabulation generally results in one-way tables which supply answers to questions about one characteristic of data only. As against this, complex tabulation usually results in two-way tables (which give information about two inter-related characteristics of data), three-way tables (giving information about three interrelated characteristics of data) or still higher order tables, also known as manifold tables, which supply information about several interrelated characteristics of data. Two-way tables, three-way tables or manifold tables are all examples of what is sometimes described as cross tabulation.

Generally Accepted Principles of Tabulation

- Every table should have a clear, concise and adequate title to make the table intelligible without reference to the text and this title should always be placed just above the body of the table.
- Every table should be given a distinct number to facilitate easy reference.
- The column headings (captions) and the row headings(stubs) of the table should be clear and brief.
- The units of measurement under each heading or sub-headings must always be indicated.
- Explanatory footnotes, if any, concerning the table should be placed directly beneath the table, along with the reference symbols used in the table.
- Source or sources from where the data in the table have been obtained must be indicated just below the table.
- Usually the columns are separated from one another by lines which make the table more readable and attractive. Lines are always drawn at the top and bottom of the table and below the captions.
- There should be thick lines to separate the data under one class from the data under another class and lines separating the sub-divisions of the class should be comparatively thin lines.
- The columns may be numerated to facilities reference.

- Those columns whose data are to be compared should be kept side by side. Similarity, percentage and/or averages must also be kept close to the data.
- It is generally considered better to approximate figures before tabulation as the same would reduce unnecessary details in the table itself.
- In order to emphasis the relative significance of certain categories, different kinds of type spacing and indentations may be used.
- It is important that all column figures be properly aligned. Decimal points and (+) or (_) signs should be in perfect alignment.
- Abbreviations should be avoided to the extent possible and ditto marks should not be used in the table.
- Miscellaneous and exceptional items, if any, should be usually placed in the last row of the table.
- Table should be made as logical, clear accurate and simple as possible. If the data happen to be very large, they should not be in a single table for that will make table unwieldy and inconvenient.
- Total of rows should normally be placed in the extreme right column and that of columns should be placed at the bottom.
- The arranged to the categories in a table may be chronological, geographical, alphabetical or according to magnitude to facilitate comparison. Above all, the table must suit the needs and requirement of an investigation.

3.6 Graphical Representation

Graphs help to understand the data easily. All statistical packages, MS Excel, and OpenOffice.org offer a wide range of graphs. In case of qualitative data (or categorised data), most common graphs are bar charts and pie charts.

Bar chart: A bar charts consist of a series of rectangles (or bar). The heights of each rectangle are determined by the frequency of that category. Suppose that the scales of a popular soft drink in the year 2010-11, in five geo graphical regions, denoted as A, B, C, D and E, are 15245,23762,9231,14980, and 12387, respectively, measured in 10000 USD. A bar chart of this data is as below.

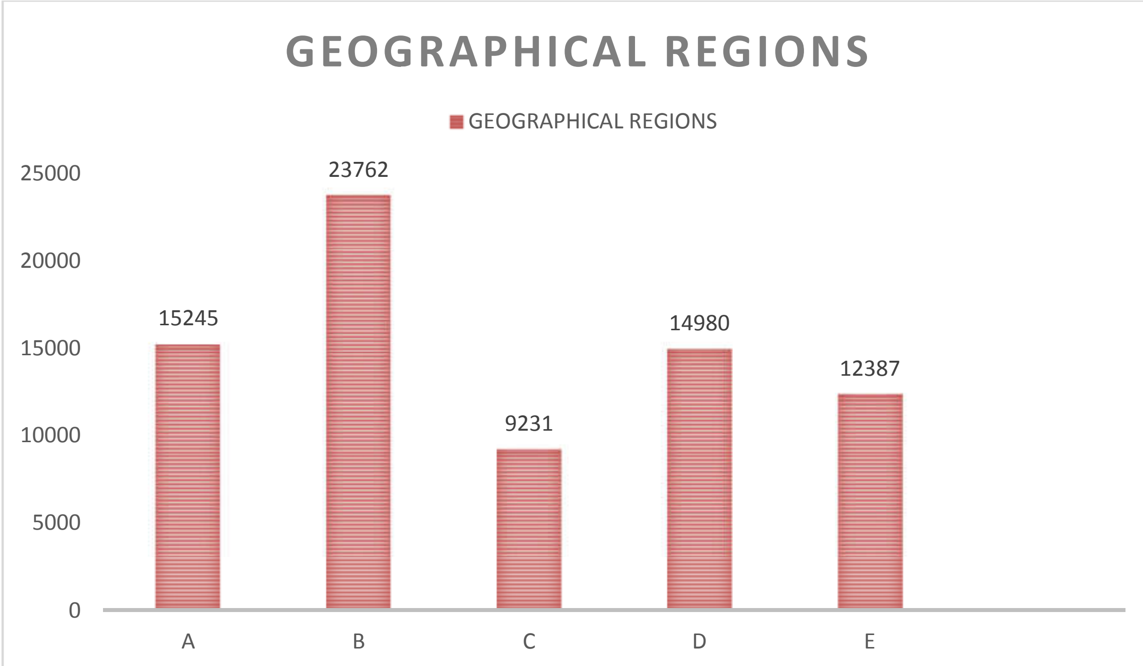


Fig. 1.1: A Bar Chart

A line chart can also be plotted in this data by connecting the midpoints of each rectangle as below:

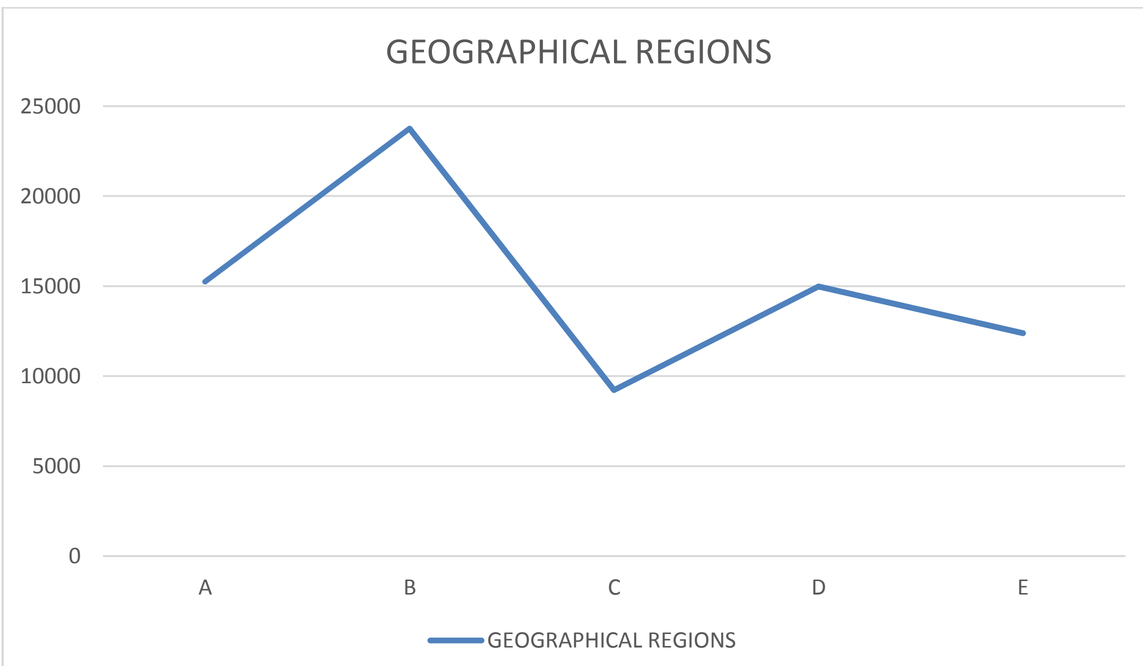
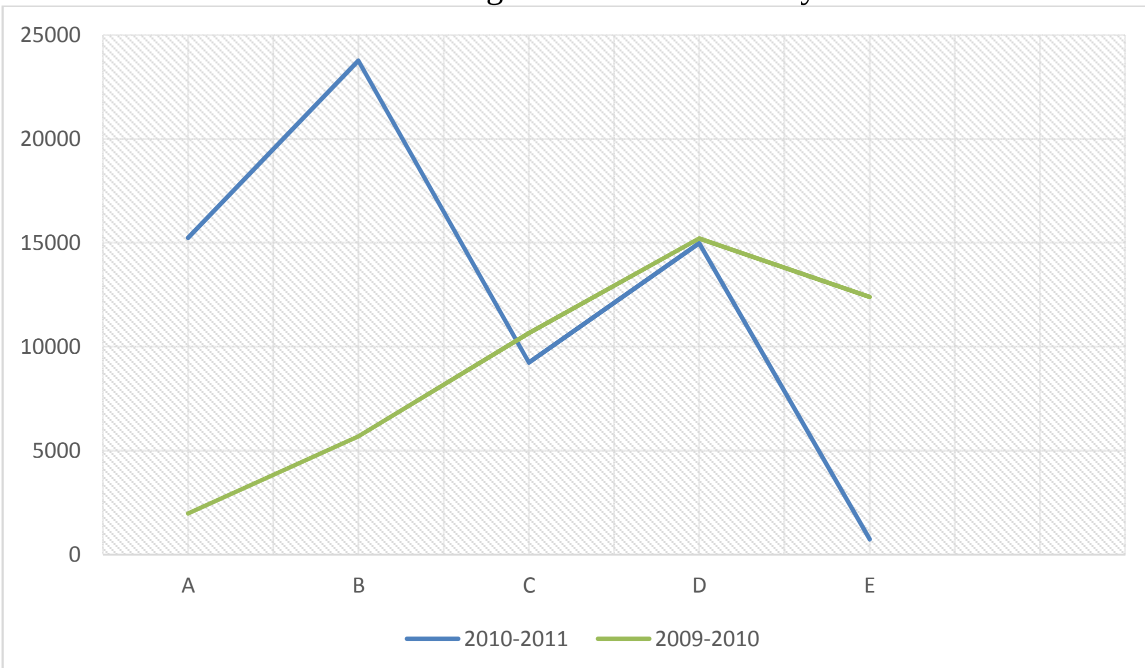


Fig.1.2: Line Charts

Line charts are useful when we wish to compare to data sets as we can overlap to line charts. For example, the scales data of the same soft drink in the same geographical regions in the same geographical regions in the year

2009-10 were 1975,5686,10652,15199 and 726, respectively, measured in 10000 USD. The line chart showing the data for both the years is



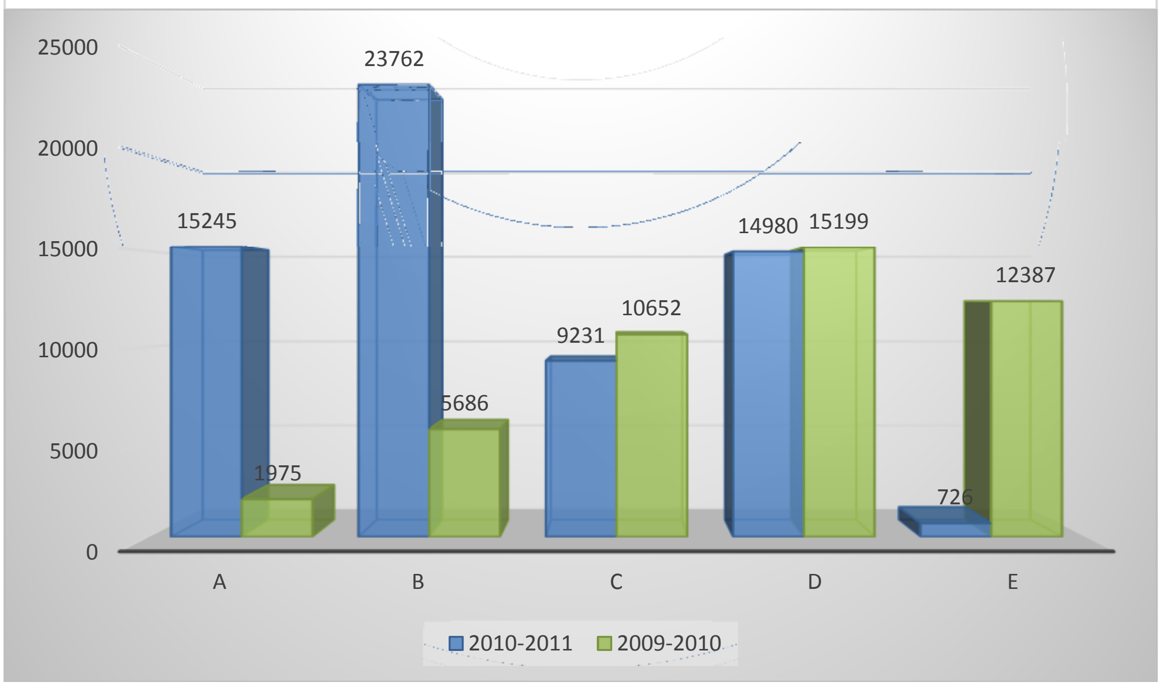


Fig.1.3:

Pie Chart: A pie chart is used to emphasize relative proportion or shares pf each category. It's a circular chart divided into sectors, illustrating relative frequencies. The relative frequency in each category or sector is proportional to the arc length of that sector or the area of that sector or the area of that sector or the central angle of that sector. Suppose in the previous example, if the soft drink has their markets only in these regions are 15245, 23762, 9231, 14980 and 12387 respectively measured in 10,000USD. A total sale of the soft drink is 75605 times 10,000USD. A pie chart can be plotted to have the idea of the shares of different markets.

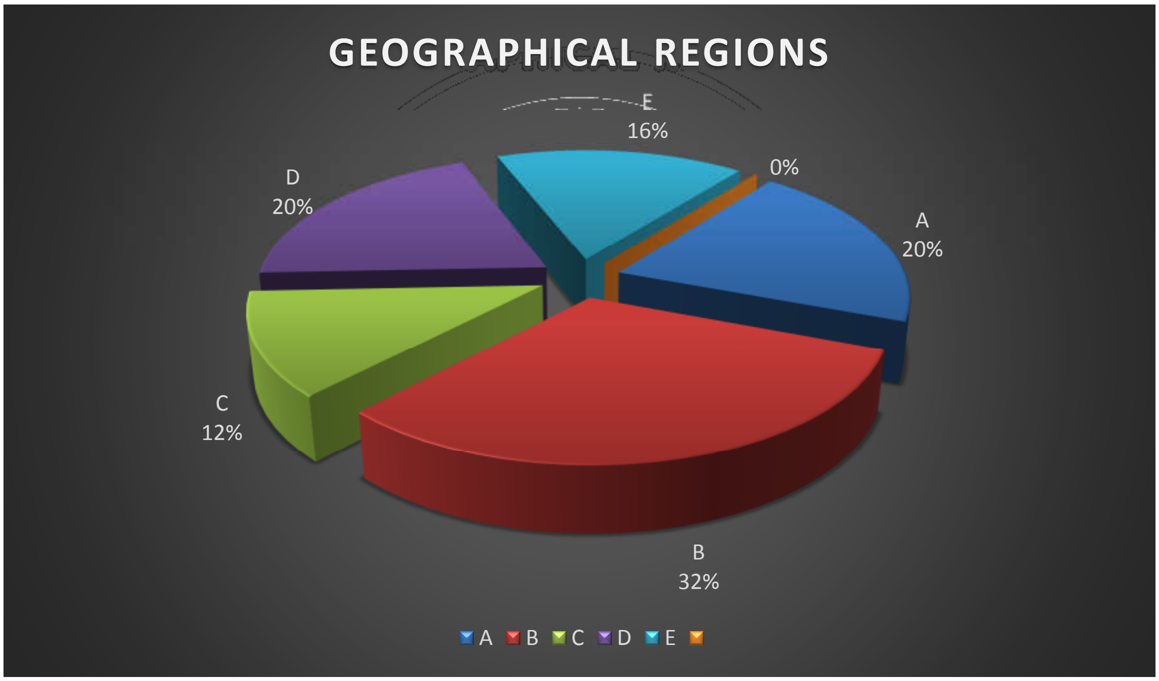


Fig.1.4: Pie Chart

In case of quantitative data, one important chart is histogram which is a generalisation of bar chart. The data is first summarised in terms of class intervals and each bar represents a class interval. The width of the bar is proportional to the width of corresponding class interval. The area of the bar is proportional to the frequency of corresponding class interval. After making the class intervals in a quantitative data set, a pie chart can also be used to read the share of each class interval.

3.7 Data Cleaning

This includes checking the data for consistency and treatment for missing value. Preliminary consistency checks are made in editing. Here we check for consistency in an extreme manner. Consistency checks look for the data which are not consistent or outlines. Such data may either be discarded or replaced by the mean value. However, the researcher should be careful while doing this. Extreme values or outlines are not always erroneous.

Missing values are the values which are unknown or not answered by the respondent. In place of such missing values some neutral value may be used. This neutral value may be the mean of available values. The other option could be to use the pattern of responses to other questions to calculate a suitable substitute to the missing values.

3.8 Data Adjusting

Data adjusting is not always necessary but it may improve the quality of analysis sometimes. This consists of the following methods;

i) **Weight assigning:** Each respondent or case is assigned a weight to reflect its importance relative to other respondents or cases.

Using this method, the collected sample can be made a stronger representative of a target population on specific characteristics. For example, the cases of educated people could be assigned higher weights and of uneducated people could be assigned lower wrights in some surveys. The value 1.0 means unweight age case.

ii) **Variable Re-specification:** This involves creating new variables or modifying existing variables. For example, if usefulness of a certain product is measured on 10-point scale, it may be reduced on a 4 point— “very useful”, “useful”, “neutral”, “not useful”. Ratio of two variables may also be taken to create a new variable.

For example, a group of people is divided into smokers and non - smokers, we can define a dummy variable taking the value”1” for smokers and “0” for non-smokers.

iii) **Scale Transformation:** Scale transformation is done to ensure the comparability with other scales or to make the data suitable for analysis. Different types of characteristics are measured on different scales. For example, attitude variables are measured on continuous scale, life style variables are usually measured on a five point Likert scale. So the variables which are measured on different scales cannot be compared. A common transformation is subtracting all the values of a characteristic by corresponding mean and dividing by corresponding standard deviation.

3.8.1 Missing Values and Outliers

Missing values are the observations which the researcher plans to collect but could not collect or lost due to some reason. Many statistical tools cannot be employed when the data set has one or more missing values. In data collection through asking question 'don't know' response may also creep-in the problem of missing values. Utmost care should be taken by the research to avoid the missing value set. Most common methods to deal with the problem of missing values while conducting the analysis is either to leave the observation, if possible, or to replace the missing values by the arithmetic mean of other collected observation.

Outliers are the observations which are quite from other observation in the data set. Although all statistical techniques can be employed when data set has outliers, their interpretation may be misleading. The most common of the outliers being present in the data set is the recording error. This error should be corrected while editing and cleaning the data. Consider an example of the survey of 100 customers in a mall.

If few bulk customers purchasing very large amount are among the 100 survey customers. In this survey having outliers (bulk customers) may not be posing any error as bulk customers are always there in the mall along with small customers. However, in a similar survey at a nearby grocery shop on a day when there is strike in the mall may include some bulk customers which could be misleading. Thus outliers should not be ignored as they might have some relevant information pose to a serious risk.

Before detecting the outliers, we need to define them first. Commonly, an observation with a value that is more than three standard deviations from the mean is considered as an outlier. A scatter plot (discussed later) can also be helpful in identifying the outliers. After identifying an outlier, the research has to decide what to do with it. The research may like to delete it or modify the value of it or retain it as it is. It depends on the knowledge about the cause of that outlier.

3.8.2 Types of Analysis

As stated earlier, by analysis we mean the computation of certain indices or measures along with searching for pattern of relationship that exists among the data groups. Analysis, particularly in case of survey or experimental data, involves estimating the values of unknown parameters of the population and testing of hypothesis for drawing inferences. Analysis may, therefore, be categorised as descriptive analysis and inferential analysis (inferential analysis is often known as statistical analysis). Descriptive analysis is largely the study of the distributions of one or more variables involved in the study. In this context we work out various measures that show the size and shape of distribution(s) along with the study of measuring relationship between two or more variables.

We may as well talk of correlation analysis and casual analysis. *Correlation analysis* studies the joint variation of two or more variables for determining the amount of correlation between two or more variables. Casual analysis is concerned with the study of how one or more variables affect changes in another variable. It is thus a study of functional relationships existing between two or more variables. This analysis can be termed as regression analysis. Casual analysis is considered relatively more important in experimental researches, whereas in most social and business researches our interest lies in understanding and controlling relationship between variables then with determining causes per se and as such we consider correlation analysis as relatively more important.

In modern times, with availability of computer facilities, there have been the rapid developments of *multivariate analysis*. Usually the following analyses are involved when we make a reference of *multivariate analysis*:

a) **Multiple regression analysis**: This analysis is adopted when the research has one dependent variable which is presumed to be a function of two or more independent variables. The objective of this analysis is to make a prediction about the dependent variable based on its covariance with all the concerned independent variables.
b) **Multiple discriminate analysis**: This analysis is appropriate when the researcher has a single dependent variable that cannot be measured, but can be classified into two or more groups on the basis of some attribute. The object of this analysis happens to be to predict an entity's possibility of belonging to a particular group based on several predictor variables.
c) **Multi analysis of variance (or multi-ANOVA)**: This analysis is an extension of two-way ANOVA, wherein the ratio of among group variance to within group variance is worked out on a set of variables.
d) **Canonical analysis**: This analysis can be used in case of both measurable and non-measurable variables for the purpose of simultaneously predicting a set of dependent variables from their joint covariance with a set of independent variables.
e) **Inferential analysis**: This is concerned with the various tests of significance for testing hypothesis in order to determine with what validity data can be said to indicate some conclusion or conclusions.

It is concerned with the estimation of population values. It is mainly on the basis of inferential analysis that the task of interpretation (i.e., the task of drawing inferences and conclusions) is performed.

4.0 CONCLUSION

The data, after collection, has to be prepared for analysis. The collected data is raw and it must be converted to the form that is suitable for the required analysis. The results of the analysis are affected a lot by the form of the data. So, proper data preparation is a must to get reliable results.

Questionnaire checking, editing, coding, classification, tabulation, graphical representation, data cleaning, and data adjusting, in like manner, missing values and outliers, various types of analysis, such as multiple regression analysis, multiple discriminate analysis**,** multivariate analysis of variance(ANOVA) and canonical analysis were discussed briefly in this unit.

5.0 SUMMARY

We have learnt the following in this unit;

- Questionnaire checking
- Editing
- Coding
- Classification
- Tabulation
- Graphical representation
- Data cleaning
- Data adjusting
- Missing values and outliers
- Various types of Analysis.

6.0 TUTOR-MARKED ASSIGNMENT

1. Write brief notes on the following methods of data preparation
    a) Questionnaire checking
    b) Editing
    c) Coding
    d) Classification
    e) Tabulation
    f) Graphical representation

2. What do you understand by the term data cleaning with respect to data preparation?
3. How would you carry out the process of data adjusting?
4. "State all the different methods under this particular process in data preparation"
5. Concisely discuss the various types of data analysis you have learnt.

7.0 REFERENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2^{nd} ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6thed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7thed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nded.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

UNIT 2 PUBLIC HEALTH STATISTICS

CONTENTS

1.0 INTRODUCTION

Routine Health Data is the use of routine vital and health statistics to describe the distribution of disease in time, place and person. Information about a defined population that is collected in consistent manner for administrative reasons is often called a routine data. These data may be used to describe the needs and services of and services provided to different population groups. The sources of routine health data, the information collected and the frequency of collection vary between different countries. However, almost all countries process their data into vital statistics. These concern the important events in human life, such as births, deaths and migrations.

Two standard and important vital statistics used across the globe for assessing a population‘s health are life expectancy and infant mortality. Life expectancy is the expectation of life at birth. It is defined as the period after which half of all persons born have died. Infant mortality is the number of children per 1000 live births who die in their first year of life.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- describe strengths and weaknesses of vital statistics
- describe epidemiology viz-time, place and person
- calculate mortality rate using person-time at risk
- summarise data through the use of qualitative data viz- nominal and ordinal terminologies,
- measure the spread of observation.

3.0 MAIN CONTENT

3.1 Strengths and Weaknesses of Vital Statistics

Table 2.1: Strengths and Weaknesses of Vital Statistics

Strength of Vital Statistics	Weaknesses of Vital Statistics
• Cheap and readily available. • Almost complete data recording • Contemporary. • Can be used for ecological studies to develop hypotheses. • Recorded at regular intervals – can be used for following trends.	• Incomplete. • Potential for bias(e.g. postmortem inflation of socio economic status: diseases with stigma under-reported) • Currency can become out of date (e.g. census data only recorded every 10 years).

Improving the reliability, validity, and completeness of routine data is important to avoid waste and minimise the use of resources. There should therefore be a good reason to begin or to stop collecting each item of data. The quality of data can be improved as shown below;

3.2 Improving Data Quality

Table 2.2:Improving Data Collection

Computerised data collation and analysis	Improves the accuracy and timeliness of the preparation and documentation of information.
Feedback	Improving feedback of collated data

	to providers is essential if their interest is to be maintained and their attention to providing quality data sustained.
Presentation	Data is to be provided in a variety of ways which are meaningful to policy makers, the media, professionals and the lay public.
Training	Training the coders and those responsible for data entry in the use of standard definitions, terminology e.t.c.

3.3 Routine Statistics: Mortality Indices

Table 2.3: Statistics Mortality Indices

Index	Typical reference period	Numerator	Denominator
Crude mortality rate	1-year	Number of deaths	Midyear population
Age specific mortality rate	1-year	Number of deaths aged X	Midyear population aged X
Child mortality rate	1-year	Number of death under 1 year old	Number of live birth
Postnatal mortality rate	1-year	Number of death in infant aged 4-52 weeks	Number of live birth
Neonatal mortality rate	1-year	Number of death in the first 28 days	Number of live birth
Perinatal mortality rate	1-year	Number of stillbirth + death <7 days	number of live births + stillbirth
Standardized mortality ratio	See sec		

Rate	**Typical reference period**	**Numerator**	**Denominator**	**Advantages/limitations**
Crude birth	1 year	Number of live births	1000 total population	Poor indicator of fertility: denominator

rate				includes man, children and postmenopausal women
Genera l fertility rate	1 year	Number of live birth	1000 women aged 15-44 years	Denominator include only women; typically of childbearing age
Age-specific fertility rate	1 year	Number of live birth in 1-year	1000 women within a particular age band, e.g., 20-24years	More precise-takes into account differences in fertility at different ages
Total period fertility rate	1 year	Sum of age-specific rate across an average women's reproductive lifetime. UK=1.66 in 2005		Enable comparison between countries over time

3.4 Descriptive Epidemiology

Epidemiology is the study of the patterns, causes and control of diseases in groups of people. In descriptive epidemiology, the three major dimensions used to describe the occurrence of disease are time, place and person.

- **Time**

Considers when the diseases occur and how it changes/has changed over time, described by

Epidemic curves: acute increase in disease frequency
Seasonal variation: cyclical patterns in disease frequency, e.g., seasonal influenza

Secular trends: trends over decades and centuries
Point events: sudden emergence of disease at a particular time.

- Place

Describe where the incidence is high/low, where incidence is change/has change, consider, considering:

- Age
- Sex
- Occupation/social class
- Ethnicity
- Behavior/life style.

3.5 Ratios (Numerators and Denominators) and Population at Risk

In epidemiology, a numerator is a feature that has been counted (e.g. number of deaths); it forms the upper part of a fraction. A denominator, for epidemiology purposes, is usually the population from which the numerator is dawn. The denominator is the lower part of a fraction and is often restricted to a particular time or period.

A population at risk is a group of a wider population that has been subject to the exposure of interest and which therefore has a propensity to have been affected by that exposure. For example, the population at risk of prostate cancer is all males who have not previously had a prostatectomy.

The numerator and denominated can be combined into a ratio, a proportion or a rate, as detailed below. Some of the most frequent mistakes in applying epidemiological principles to real-world problems come from a failure to define and to count numerators and denominators accurately.

Ratio

Ratio=$n_1/n_2=n_1:n_2$

Where n_1 and n_2 are numbers.

For example,

Ratio of males to females in a population= $\frac{\text{number of males in a population}}{\text{number of females in a population}}$

A ratio is often expressed as odds, which is a single number. For example, if a bag contains two white balls (n_1), three black balls (n_2) and five grey balls (n), then the ratio of black balls to white balls is 3:2 (three to two). This can be simplified by dividing both the numerator and the denominator by the denominator to give 1:5:1 (one and a half to one). Expressed in odds, it is simply 1½.

Proportion

Proportion n_1/N

Where n_1 is a subpopulation of the whole study population, N. if a bag contains two white balls (n_1), three black balls (n_2) and five grey balls (n_3), then the proportion of black balls is 3/10 or 0.3.

For example,

Proportion of men in a population=$\frac{\text{Number of males in a population}}{\text{Number of males+Number of females in a population}}$

Rate

Rate=$\frac{N}{P \times T}$

Where N is the numerator, P is the number of people in the population and T is a period of time. This is a measure of frequency of the occurrence of a phenomenon. The denominator is constituted from both population and time.

For example,

$$\text{Incidence of epilepsy} = \frac{\text{Number of new cases of epilepsy}}{\text{Population x Reporting period}}$$

Denominators should include only those at risk, and not those who cannot possibly develop the disease. For example, people who have been immunised against a disease should not be included in studies looking at the rate at which people acquire that disease, since they cannot be considered in the same way to be at risk at those who have not been immunised.

Time at Risk

Time at risk describes the total amount of time that individuals within a study spend at risk of developing the disease of interest. The concept is particularly important in the analysis of cohort studies. One of the problems with conducting cohort studies is that some subjects will join the study after data collection has started. Others will leave the at-risk population early because they have:

- Died
- Moved away
- Become a case
- Been lost to follow-up for another reason (e.g. withdrawn from the study)
- Been censored

To correct for this variation, the 'total person-time at risk' is used as the denominator in calculations of morbidity or mortality relating to cohort studies. Person-time represents the sum of all the individual's times at risk.

$$\text{Incidence rate} = \frac{\text{Number of new cases in specific period}}{\text{Total person-time at risk}}$$

The box below represents a cohort study, and illustrates the concept of time at risk. As time progresses, the number of individuals at risk will fall as people die, become a case or at lost to follow-up. Censoring occurs when the value of an observation is only partially known. At the end of this study, patient 3 was still alive and therefore spent at least 10 days at risk. This is known as right-censoring. Left-censoring (where a data point is below a certain value but is known by how much) and interval censoring (where a data point is somewhere in the interval between two values) are also possible.

Sometimes, participants are replaced by new people recruited into the study. For studies of short duration and where few individuals leave or join the at-risk population, it is reasonably accurate to use the number of individuals at the start of the study as the denominator in incidence calculations.

Table 2.2: Cohort Study Illustrating the Concept of Time at Risk

	Patient 1	Patient 2	Patient 3	Patient 4
Date of entry to study	1998	2000	1995	1997
Event	Death	Loss	Censored	Death
Date of event	2004	2003	2005	1997
Person-years in study	6	3	10	2

3.6 Calculation of Mortality Rate Using Person-Time at Risk

Where the study duration is longer or the likelihood of individuals leaving the risk at-risk population is greater, it is preferable to use a more accurate calculation of the incidence rate using Person-time at risk. See table below

Table 2.3: Person-Time at Risk

Calendar time	Person-years	Events	Death rate (number/year)
1995-1999	8	1	0.125
2000-2004	12.5	1	0.080
(2005-2009)	(0.5)	(0)	(0)

3.7 Methods for Summarising(Qualitative) Data

Data are considered as either qualitative (non-numeric) or quantitative (numeric). Subtypes of each type of data are listed below.

Qualitative Data

Qualitative data are non-numeric. Some qualitative data are categorical: they describe different categories or states that a subject may fall into. These may be further categorised into nominal data and ordinal data.

Nominal Data
Nominal data have no order and thus only give names or labels to various categories. Examples include the ABO blood groups (A, B, AB, O), the names of colours (red, yellow, orange, green, violet, purple) and types of hospital ward (rehabilitation, surgical, medical)

Ordinal Data
Ordinal data have order, but the interval between measurements is not meaningful.

Some qualitative data are ordinal data: there is a natural order to the states, but no clear numerical relationship between them. Examples are [poor, fair, good, better, best] and [very quiet, quiet, normal value, loud, loudest].

Although ordinal data should not be used for calculations, it is not uncommon to find averages calculated of the type strongly disagree, disagree, neither agree nor disagree, agree, strongly agree.

3.8 Quantitative Data

Quantitative data are numeric, and they are further classified as either discrete or continuous. Data may come from one or more of the following categories.

Discrete Data
Discrete data have a finite number of possible numerical values. Examples include the number of children with brown eyes in a class of 30 children, or the number of times someone has been admitted to hospital in their lifetime.

Continuous Data
Continuous data include measurable quantities of length, volume, time, mass, etc. they frequently have an upper or lower limit, e.g. height cannot be <0.

Interval Data
Interval data have meaningful intervals between measurements, e.g. the age groups 0-4, 5-9, 10-14…….90+. They are typically displayed as a table or histogram.

Ratio Data
Ratio data are the most flexible data to work with, since they contain the most information of any data type. It becomes meaningful to say not only

that A scored 1 and B scored 2, but that B is twice as good as A. Ratio data are ideal for use as outcome variables in regression.

Binary Data

Binary data are a special type of data that has just 2 values. Depending on how they are analysed, they may be considered to have properties of ordinal data, interval data or ratio data. They are very common data in epidemiology, since they accurately describe many of the states that are of interest. For example, did a patient improve following a particular treatment – or not? Is the case alive or dead? Was the case in the treatment group or the control group?

Summarising Data

Discrete variables (e.g. blood groups) are typically summarised as proportions. For example, the distribution of the common ABO blood groups in England is shown in the table below.

Continuous variables (e.g. blood pressure) are described using both:

- A measure of central tendency (mean/median/mode)
- A measure of spread (range/variance/standard deviation).

If a population is normally distributed, then it can be described by both its, mean and standard deviation alone.

Prevalence of different ABO blood groups in the UK population

Table 2.4: Blood Groups

Blood group	Proportion
A	0.41
B	0.08
AB	0.03
O	0.48

3.9 Measures of the Spread of Observations

Table 2.5: Some Commonly Used Terms are Listed in the Table

Measure	**Description**
Range	The difference between the largest and smallest value
Percentiles	The value below which p% of the observations in a population fall is called the pth percentile
Interquartile range	The difference between the value at the 25th centile and the value at the 75th centile (i.e. between the 1st and 3rd quartiles)

Standard deviation	Measure of the spread of observations about the mean of the samples. Takes the same units as the data; 1.96 standard deviations either side of the mean covers 95% of the population Used to describe the data ‘Deviation – Description’ Square root of the variance of the sample
Standard error of the mean	The error associated with measuring a sample mean is determined by both the standard deviation and the number of observations in the sample. The more observations, the more precise is the estimate of the mean Measure variability of the mean of the sample Used to make estimations about the true mean of the population
Central limit theorem	If repeated samples are taken from any population, then the means of these samples will tend towards a normal distribution, even if the population is not normally distributed

4.0 CONCLUSION

Routine Health Data is the use of routine vital and health statistics to describe the distribution of disease in time, place and person. Information about a defined population that is collected in consistent manner for administrative reasons is often called a routine data. These data may be used to describe the needs and services of and services provided to different population groups. Two standard and important vital statistics used across the globe for assessing a population‘s health are life expectancy and infant mortality.

5.0 SUMMARY

In this unit we have learnt about the following:

- Strengths and weaknesses of Vita statistics
- Improving Data Quality
- Descriptive Epidemiology viz-Time, Place and Person
- Ratios and Populations at Risk
- Calculation of Mortality rate using Person-Time at Risk
- Methods of Summarising Data through the use of Qualitative Data viz- Nominal and Ordinal terminologies,
- Likewise, Quantitative data viz- Continuous, Interval, Ratio and Binary
- Measure of the Spread of Observation

6.0 TUTOR-MARKED ASSIGNMENT

1. In tabular form only show the strengths and weaknesses of vita and health statistics.
2. What do you understand by descriptive epidemiology?
3. Explain the terms Qualitative and Quantitative Data as they are applicable in health and vital statistics.
4. In 1980, the mid- year population of a town X was 700 million and the total deaths were 79.4 million. Calculate the death rate for this year.
5. There were 40 motor accidents in Y town in Nigeria with a population of 150,000 in the year 1987. Calculate the incidence rate of accidents.

7.0 REFERENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nded.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim&Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6thed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7thed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nded.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

MODULE 4

UNIT 1 FURTHER FEATURES OF MORBIDITY AND MORTALITY INDICES

CONTENTS

1.0 INTRODUCTION

Incidence and prevalence (direct and indirect standardization)
Incidence and prevalence are measures of occurrence.

Incidence relates to new occurrence. There are three related measures of incidence as shown in the table below.

Prevalence relates to existing outcomes. Prevalence is also called the 'point prevalence', i.e. the proportion of a population with a disease. It is approximately equal to incidence x duration provided that the incidence and the death/recovery ate have been stable for the disease over some preceding time. Period prevalence relates to the proportion of the population with a disease during a specified period. These two measures of prevalence are shown in the table below.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- explain measures of incidence and prevalence
- discuss direct standardisation
- measures disease burden.

3.0 MAIN CONTENT

3.1 Measures of incidence

Table 1.1: Measures of incidence

Concept	Alternative name	Definition
Incidence	Incidence rate	Number of new events during a specified time period
Incidence rate	Force of morbidity Incidence density	Number of new events divided by the total person-time at risk Used to describe studies where there have been varying periods of follow-up (e.g. 3 cases during 25 person years = 12 cases per 100 person-years)
Cumulative incidence	Risk	Proportion of a population who become diseased in a defined time period. This is a measure of the risk that an individual will become diseased during a defined time period, e.g. the attach rate during an epidemic

Note:

1. Incidence (unlike prevalence) is not affected by disease survival
2. The denominator should include only those ‘at risk’

Rather confusingly, the term incidence rate is sometimes used to mean incidence. The incidence and incidence rate can generally be considered as the same concept, with the number of person-years being more or less accurately measured.

3.2 Measures of Prevalence

Table 1.2: Measures of Prevalence

Concept	Definition
Point prevalence	Proportion of a population with a disease at a particular point in time
Period prevalence	Proportion of a population who had the disease during a specified period

Note: since period prevalence considers both cases at the start and new cases arising during the time period, it includes features of both prevalence and incidence.

Standardisation

Also known as adjustment, this technique is required to make comparisons between populations of differing demographic structures, where crude mortality rates would be misleading. For example, the technique is typically the first step in addressing questions such as, 'Does town A have a higher mortality rate than town B?' or, 'Are 10 incident cancers in a workforce of 100 people over a 10-year period more or less than we would have expected had national rates applied?' the technique is mostly used to compare populations that differ in their age structure – but it may also be applied to correct for differences in gender or social class (or combinations of these and other variables).

To standardise data, populations are divided into strata based on differences that might potentially influence the comparison that is being made. For example, to compare the numbers of deaths in a retirement community with a town having a high proportion of young mothers, the mortality rates would need to be age standardised.

Two forms of standardisation that are commonly used are direct and indirect

3.3 Direct Standardisation

This may be used to compare two populations in different regions, or the same population at two different periods of time. In direct standardisation, the age- or stratum specific mortality rates of the observed population are known. (E.g. rate of lung cancer in 20 to 40-year-old men). These rates are then applied to standard or reference population. Data typically need to be available concerning a large number of subjects in order to have adequate numbers in each stratum to be confident about the estimate.

- Start by choosing a single population (e.g. one of those being compared, their average or an outside population).
- Break this single population being compared, take the age-specific mortality rate for each age band and multiply it by the size-weighting of that age band from the standard age structure
- Sum of all these to obtain the age-standardised mortality rate. The actual value of this adjusted rate is meaningless but it shows the true difference in rates between the populations being compared.

The European Standard Population

This is used to compute directly age-standardised rates. The sample population is used for males, females and all persons.

The example in the table below shows how mortality data from twoyears can be compared using direct standardisation.

Table 1.3: Mortality Data

Age Group	**Number in Age Group**
0	1600
1-4	6400
5-9	7000
10-14	7000
15-19	7000
20-24	7000
25-29	7000
30-34	7000
35-39	7000
40-44	7000
45-49	7000
50-54	7000
55-59	6000
60-64	5000
65-69	4000
70-74	3000
75-79	2000
80-84	1000
85+	1000
TOTAL	100 000

3.4 Indirect Standardisation

With indirect standardisation, the population under study is usually small. The technique hinges on the calculation of the rate of death that would have been expected in the study population had a comparison rate applied instead. This permits the determination of the standardised mortality ratio (SMR).

- Start with the stratum-specific death rates of a standard population (e.g. European Standard Population if age is the effect to be standardised)
- Use these to calculate expected number of deaths in each stratum of the population
- Add up the expected number of deaths for each age band:
  Define E as the expected number of deaths
  Define K as the number of strata
  Define n_i as the number of people in the ith stratum
  Define R_i as the rate of death in the ith stratum
- Standardised mortality ratio (SMR) is then calculated as (Observed deaths/Expected deaths)

The SMR indicates whether mortality in the study group (after correction for age, sex, etc.) is unusual when compared with standard population. In occupational mortality studies, comparisons are often made against two standard populations: (i) an unexposed population from the same occupation and (ii) the general population.

When used to estimate the association between an occupational exposure and a disease, the SMR underestimates the true magnitude of association because the general population contains both exposed and unexposed individuals.

3.5 Years of Life Lost

One way of considering the impact of a particular disease or risk factor is to consider how many years' people might have expected to have lived had their lives not been curtailed by the disease. For example, deaths from road traffic collisions affect mostly young males who otherwise may have been expected to live their 70s or 80s.

'Years of life lost (YLL)' is a measure of premature mortality and it explicitly places more importance upon deaths that occur in the young than upon those in the elderly. It is therefore a value-laden statistic, which reflects the wish of society to prevent avoidable causes of death in younger people, and to avoid the loss to society of investment in raising children and young adults.

To calculate YLL in its most simple form, an upper age limit (e.g. 75) is chosen. A person dying at the age of 60 contributes 15 YLL (75 – 60 =15) to the calculation. A person dying at age 15 contributes 60 years. Deaths occurring in people aged over 75 are not included in the calculation. Infant deaths may or may not be included.

In more complex situations, an actuarial assessment of the expectation of life for each person will be undertaken. Life-expectancy is calculated using current life-tables: age-specific mortality rates are applied to a hypothetical population.

3.6 Measures of Disease Burden

The measures of disease burden (event and time based) and population attributable risks, including identification of comparison groups appropriate to public health.

A number measures exist to assess the weight of disease affecting a community, taking into account both the number of events and the impact of these events on the population. They each depend on the availability of data.

Measures of Disease Burden Based on Events
These studies of incidence require data from routine sources:

- Death certificates (if the disease burden is in mortality rather than morbidity)
- Hospital episode data (if the disease results in hospital admissions)
- Disease register (if dedicated register for that data exists)
- Statutory notifications: infectious diseases, procedures, e.g. abortion.

Measures of Disease Burden
Where there is no routine event-based data collection, a prevalence ('cross-sectional') survey may be needed.

Variation

Sources of variation, its measurements and control
Variation arises from difference between populations or between individuals within a population. It may be due to random or non-random factors.

Measurement can differ for many reasons. Some differences will be due to genuine dissimilarities between subjects. This sort of variation is often random, and follows a normal distribution. Some people are taller than average, some are shorter than average. This is every day, random variation.

Other differences will be due to discrepancies the way that measurements were made or recorded. A ruler may measure only to the nearest 5cm. it judges a person's height as 175cm and another's as 170cm. A more precise ruler, accurate to 0.1cm, would confirm that the first subject is in fact 176.1cm tall and the second is 172.4cm tall. The aim for any investigator is to minimize the error associated with making the measurements, thereby focusing attention on the real differences between individuals.

Measurements can be biased or unbiased, precise or imprecise, or some combination of these.

By way of analogy, consider a markswoman firing shot at a target, her sights might be on-target or off-target, and she might be a good shot or bad shot.

3.7 Errors in Epidemiological Measurement

There are common errors in epidemiological measurements, their effect on numerator and denominator data, and their avoidance.

Measurement error can be defined as any mistake that occurs during the process of applying a standard set of values (i.e. a measurement scale) to a set of observations. In epidemiology, such errors can lead to the misclassification of cases and controls. Acknowledgment of measurement error is important in research, and leads to more robust and defensible scientific results.

Measurement errors are neither random (i.e. occurring due to chance) or systematic (i.e. persistent, non-random, differences between the observed measurement and its true value). These errors will lead respectively to non-differential errors (affecting all groups equally) and differential errors (affecting one group more than another). The table below compares the random and systematic errors in epidemiological measurement

Table 1.4: Random and Systematic Error in Epidemiological Measurement

Random error	**Systematic error**
Occurs due to chance	Occurs due to non-random factors
Affects all groups equally (non-differential)	Affects some groups more than others (differential)
Exposure: misclassification of exposures is equal for cases and controls (e.g. 20% of cases and 20% of controls are misclassified as being exposed)	Exposure: misclassification of exposures differs between cases and controls (e.g. 20% of cases misclassified as having been exposed as having the disease)
Outcome: misclassification of outcomes is equal foe exposed and non-exposed (e.g. 30% of exposed and 30% of non-exposed are misclassified as having the disease)	Outcome: misclassification of outcomes differs between exposed and non-exposed (e.g. 19% of the exposed are misclassified as having the disease but only 7% of non-exposed are misclassified as having the disease)
Bias is towards the null hypothesis (i.e. dilution of the study's findings)	Bias can be in aby direction
Less threatening for a study than systematic error	More threatening for a study than random error
Example: a study of lung cancer in relation to proximity of residence to a coke oven classifies subjects (cases and population) by distance of residence from the oven at the time of follow-up. Here there is misclassification due to migration (not all people living near the oven at the time of follow-up will have lived there at the etiologically relevant time) but this error occurs randomly	Example: a study assessing the association between visual display unit (VDU) usage and spontaneous abortion. Here, cases are more likely to recall VDU usage compared with controls, particularly if there has been media interest in this hypothesis. Therefore the association measured is likely to be greater than the true association

Effects of Measurement Errors

Measurement errors can affect:

- Dependent variables (outcomes such as disease)
- Independent variables (risk factors such as exposures)
- Confounder variables
- Effect modifiers

With regard to the measurement of confounder variables, an apparent association between the exposure and the disease may persist after statistical adjustments. Such residual confounding is particularly problematic where the variable being measured is difficult to quantify (e.g. measurement of socioeconomic status). For this reason, residual confounding should always be considered where an association persists following statistical adjustment for a known confounder

3.8 Avoidance of Measurement Error

Measurement error can be avoided or accounted for at various stages of the study.

Table 1.5:Avoiding and Accounting for Measurement Error

Type of study	Strategy for dealing with measurement error
Design	Set out to measure reliability: correlation coefficients (continuous variables) or Cohen's kappa (categorical variables) Blinding Standardised measurement instruments Use multiple sources of information (questionnaires, direct measurements, registries, case records) Use multiple controls
Data collection	Administer instruments equally to: Cases/controls Exposed/unexposed
Data analysis	Perform a sensitivity analysis to test the robustness of the findings
Reporting	Consider and mention potential random and systematic errors

4.0 CONCLUSION

Incidence relates to new occurrence, while Prevalence relates to existing outcomes. Prevalence is also called the 'point prevalence', i.e. the proportion of a population with a disease. It is approximately equal to incidence x duration provided that the incidence and the death/recovery rate have been stable for the disease over some preceding time. Period prevalence relates to the proportion of the population with a disease during a specified period. Incidence and prevalence are both measures of occurrence.

5.0 SUMMARY

In this unit you have learnt the following:

- Measures of incidence
- Measures of Prevalence
- Direct Standardisation
- Indirect standardisation
- Years of life lost
- Measures of disease burden
- Errors in Epidemiological measurement
- Avoidance of measurement errors.

6.0 TUTOR-MARKED ASSIGNMENT

1 Define the following:
  a) Incidence rate
  b) Prevalence rate
  c) Point prevalence rate
  d) Period prevalence rate
  e) Years of life lost.

2 In a Hamlet with a population of 6000, 30 new cases of Diarrhea occurred in February 1976 and 52 cases in 1977, in the months of July, August, and September. Calculate monthly morbidity rate in 1976 and 1977.

7.0 REFRENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6^{th} ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7^{th} ed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2^{nd} ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

UNIT 2 SUMMARY OF BIOSTATISTICS

CONTENTS

1.0 INTRODUCTION

Biostatistics is the term used when tools of statistics are applied to the data that is derived from biological sciences such as medicine.Any science demands precision for its development, and so does medical science. For precision, facts, observations or measurements have to be expressed in figures.

Everything in medicine be it research, diagnosis or treatment, depends on counting or measurement. High or low blood pressure has no meaning, unless it is expressed in figures. Incidence of tuberculosis or death rate in typhoid is stated in figures. Enlargement of spleen is measured in fingers' breath. Thus medical statistics or biostatistics can be called **quantitative medicine.**

In nature, blood pressure, pulse rate, action of a drug or any other measurement or counting varies not only from person to person but also from group to group. The extent of this variability in an attribute or a character, whether it is by chance, i.e., biological or normal, is learnt by studying statistics as a science.

Comparison of two or more groups is of great importance in applied scientific practice of medicine, e.g., infant mortality rate in developing countries like India was around 73 per thousand live births in 1994 while in developed countries like the USA, UK and Japan, the rates have gone down to about five per thousand live births per year due to external factors like socio-economic advancement, better application of scientific knowledge in medicine or improved health services. Rise in pulse rate noted after an injection of a drug may be by chance or due to the effect of drug.

Variation more than natural limits may be pathological i.e., abnormal due to the play of certain external factors. Hence, biostatistics may also be called a **science of variation**. The data after collection, lying in a haphazard mass are of no use, unless they are properly sorted, presented, compared, analysed and interpreted. They mean something more than figures, give a dimension to the problem and even suggest the solution. For such a study of figures, one has to apply certain mathematical techniques called **statistical methods**, such as calculation of standard deviation, standard error and preparation of a life table. Though these methods are quite simple and general in application, medicos follow them only when they are put in a familiar way giving day-to-day medical examples. Moreover, medical statistics merit special attention as they deal with human beings and not with material objects or lower animals. Medical observer has to give his opinion or make an impression after applying these methods.

A medical student should not depend on a statistician for the statistical analysis. For professional interpretation of his results, he should learn the application of methods himself which do not require knowledge of mathematics higher than what he or she had acquired the application of methods himself which do not require knowledge of mathematics higher than what he or she had acquired at school. However, he or she should take the guidance of a qualified statistician right from the beginning of any scientific study till drawing the conclusions. Medical Statistics go under different names when applied in different fields such as:

Health statistics in public health or community health
Medical statistics in medicine related to the study of defect, injury, disease, efficacy of drug, serum and line of treatment, etc.

Vital statistics in demography pertaining to vital events of births, marriages and deaths, these terms are overlapping and not exclusive to each other.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- Apply and use biostatistics as figures
- explain common statistical terms

- measure location-averages and percentiles
- grouped series
- apply and use percentiles.

3.0 MAIN CONTENT

3.1 Application and Uses of Biostatistics as a Science

1. In Physiology and Anatomy

i. To define what is normal or healthy in a population and to find limits of normality in variables such as weight and pulse rate ─ the mean pulse rate is 72 per minute but up to what limits it may be normal on either side of mean has to be established with certain appropriate techniques.

ii. To find the difference between means and proportions of normal at two places or in different periods. The mean height of boys in Gujarat is less than the mean height in Punjab. Whether this difference is due to chance or a natural variation or because of some other factors such as better nutrition playing a part has to be decided.

To find the correlation between two variables X and Y such as height and weight ─ whether weight increases or decreases proportionately with height and if so by how much, has to be found.

2. In Pharmacology

i. To find the action of drug ─ a drug is given to animals or humans to see whether the changes produced are due to the drug or by chance.

ii. To compare the action of two different drugs or two successive dosages of the same drug.

iii. To find the relative potency of a new drug with respect to a standard drug.

3. In Medicine

i. To compare the efficacy of a particular drug, operation or line of treatment ─ for this, the percentage cured, relieved or died in the experiment and control groups, is compared and difference due to chance or otherwise is found by applying statistical techniques.

ii. To find an association between two attributes such as cancer and smoking or filariasis and social class ─ an appropriate test is applied for this purpose.
iii. To identify signs and symptoms of a disease or syndrome. Cough in typhoid is found by chance and fever is found in almost every case. The proportional incidence of one symptom or another indicates whether it is a characteristic feature of the disease or not.

4. In Community Medicine and Public Health

i. To test usefulness of sera and vaccines in the field ─ percentage of attacks or deaths among the vaccinated subjects is compared with that among the unvaccinated ones to find whether the difference observed is statistically significant.
ii. In epidemiological studies ─ the role of causative factors is statistically tested. Deficiency of iodine as an important cause of goitre in a community is confirmed only after comparing the incidence of goitre cases before and after giving iodized salt.

In public health, the measure adopted is evaluated. Lowering of morbidity rate in typhoid after pasteurisation of milk may be attributed to clean supply of milk, if it is statistically proved. Fall in birth rate may be the result of family planning methods adopted under National Family Welfare programme or due to rise in living standards, increasing awareness and higher age of marriage.

Thus, by learning the methods in biostatistics, a student learns to evaluate articles published in medical journals or papers read in medical conferences. He understands the basic methods of observation in his clinical practice or research.

3.2 Application and Uses of Biostatistics as Figures

Health and vital statistics are essential tools in demography, public health, medical practice and community services. Recording of vital events in birth and death registers and diseases in hospitals is like book keeping of the community, describing the incidence or prevalence of diseases, defects or deaths in a defined population. Such events properly recorded form the eyes and ears of a public health or medical administrator. Otherwise it would be like sailing in a ship without compass. Thus, biostatistics as a science of figures will tell:

a. What are the leading causes of death?
b. What are the important causes of sickness?
c. Whether a particular disease is rising or falling in severity and prevalence?
d. Which age group, sex, social class of people, profession or place is affected the most?
e. The levels or standards of health reached.
f. Age and sex composition of population in a community.
g. Whether a particular population is rising, falling, ageing or ailing?
h. Which health programme should be given priority and what will be the requirements for the same?

3.3 Common Statistical Terms

One should remember before learning the methods in biostatistics, some terms used, their symbols and notations and refer as and when needed.

1. **Variable:** A characteristic that takes on different values in different persons, places or things. A quantity that varies within limits such as height, weight, blood pressure, age, etc. It is denoted as X and notation for orderly series as X_1, X_2, X_3, ...X_n. The suffix *n* is symbol for number in the series. $\sum$ (sigma) stands for summation of results or observation.
2. **Constant:** Quantities that do not vary such as $\pi = 3.142$, $e = 2.718$. They do not require statistical study. In biostatistics, mean, standard deviation, standard error, correlation coefficient and proportion of a particular population are considered as constant.
3. **Observation:** An event and its measurements such as blood pressure (event) and 120mm of Hg (measurement).
4. **Observational Unit:** The source that gives observations such as object, person, etc. In medical statistics the term individuals or subjects is used more often.
5. **Data:** A set of values recorded on one or more observational units.
6. **Population:** It is an entire group of people or study elements ─ persons, things or measurements for which we have an interest at a particular time. Populations are determined by our sphere of interest. It may be finite or infinite. If a population consists of fixed number of values, it is said to be finite. If a population consists of an endless succession of values, the population is an infinite one. It has to be fully defined such as all human beings, all families joint or nuclear and all women of 15-45 years of age or only married women, all patients, all doctors in service or in practice and so on. Such a population invariably gives qualitative data. If it is finite or limited

in number, it can easily be counted. A statistical population may also be birth weights, haemoglobin levels, readings of a thermometer, number of RBCs in the human body, etc. Such a population mostly gives quantitative data. It is finite or small in number that cannot be easily counted.

7. **Sampling Unit:** Each member of a population.
8. **Sample:** It may be defined as a part of a population. It is a group of sampling units that form part of a population, generally selected so as to be representative of the population whose variables are under study. There are many kinds of sample that can be selected from a population. Various methods employed are described later in the book.
9. **Parameter:** It is a summary value or constant of a variable that describes the population such as mean, variance, correlation coefficient, proportion, etc. Familiar examples are mean height, birth rate, morbidity and mortality rates, etc.
10. **Statistics:** It is a summary value that describes the sample such as its mean, standard deviation, standard error, correlation coefficient, proportion, etc. This value is calculated from the sample and is often applied to population but may or may not be a valid estimate of population. Though not desirable, parameter and statistic are often used as synonyms.
11. **Parametric test:** It is one in which population constants as described above are used as mean, variances, etc. and data tend to follow one assumed or established distribution such as Normal, Binomial, Poisson, etc.
12. **Non-parametric tests:** Tests such as X^2 test, in which no constant of population is used. Data do not follow any specific distribution and no assumptions are made in nonparametric tests, e.g. to classify good, better and best you allocate arbitrary numbers or marks to each category.

Notations for a Population and Sample Values

Roman letters are used for statistics of samples and Greek for parameters of population.

Table 2.1 Common Notations

Summary Value	**Sample Statistics**	**Population Parameters**
Mean		μ
Standard deviation	S	
Variance	S^2	
Proportion	P	P

Complement of proportion	Q	Q

Table 2.2: Other Symbols Commonly Used

=	Equal to
>	Greater than
<	Less than
Z	The number of standard deviations from the mean or standard normal deviate/variate
%	Per cent
Gamma	Pearson's correlation coefficient
Rho	Spearman's correlation coefficient
O	Observed number
E	Expected number
d.f. or f	Degrees of freedom
K	Number of groups or classes
P	Probability

3.4 Measures of Location- Averages and Percentiles

Measures of Central Tendency- Averages

Information from series of observations presented by rank in a sample or frequency distribution table is summarised by an observer to get answers to the following three questions:

1. What is its average or central value?
2. How are the other values dispersed around this value or what is the degree of scatter?
3. What is the shape of the distribution - is it normal?
4. Average value of a characteristic is the one central value around which all other observations are dispersed. In any large series, nearly 50 per cent observations lie above while the remaining 50per cent lie below the central value. It indicates how the values lie near the centre. In other words, it is a measure of central tendency or concentration of all other observations around the central value. Thus an average value helps:

Firstly, to find most of the normal observations lie close to the central value, while few of the too large or too small lie far away at both ends.

Secondly, to find which group is better off by comparing the average of one group with that of the other, e.g., one finds the average incubation period of cholera is smaller than that of typhoid; income of pleaders is higher than that of doctors; average daily attendance of one hospital is higher than that of another; and so on. After finding the difference, one may reason out why in one group it is more than that in the other.

Average is a general term which describes the centre of a series. There are three common types of averages or measures of central position or central tendency- mean, median and mode. They are summary indices describing the 'central point or the most characteristic value' of a set of measurements.

Mean

This measure implies arithmetic average or arithmetic mean which is obtained by summing up all the observation and dividing the total by the number of observations. Example, Erythrocyte sedimentation rates (ESRs) of seven subjects are 7,5,3,4,6,4,5. Calculate the mean.

Mean =7+5+3+4+6+4+5=34=4.86 It is the one central value, most commonly used in statistical methods.

Median

When all the observations of a variable are arranged in either ascending or descending order, the middle observation is known as median. It implies the mid value of series. ESRs of seven subject are arranged in ascending order _3,4,4,(5) is the median is almost equal to the mean. Consider another example of observations in absenteeism of school children in the series 4,6,8(10), 12,14,32. Mean 86/7=12.3
Median value=10

In this case, mean value gives a distorted result as one observation 32 is too large, so the mean as a measure of central tendency should not be considered appropriate. To have a better idea of average, one should ignore unduly high observations such as 32 in the above example. Mean of the remaining observations will be 54/6 = 9.0 which is much closer to the median, i.e. 10 than the mean 12.3 calculated with seven observations.

Median, therefore, is a better indicator of central value when one or more of the lowest or the highest observations are wide apart or not so evenly distributed, e.g.

	YEAR	
	1974	1977
Mean age at marriage	16.40	15.70
Median age at marriage	15.70	16.00

As per mean, the age at marriage in 1977, has decreased while as per median it has increased which is nearer the truth.

Another good example is the duration of stay in a hospital in general or in a specific disease ward, or disease. In such cases median is a better indicator because the stay may be unduly long in some cases.

Mode

This is the most frequently occurring observation in a series i.e., the most common or most fashionable, such as 8 mm in tuberculin test of 10 boys given below:

3,5,7,7,8,8,9,10,11,12

Mode is rarely used in medical studies. Out of the three measures of central tendency, mean is better and utilised more often because it uses all the observations in the data and is further used in the tests of significance.

Calculation of Mean

A series of observations is indicated by the letter X and individual observations by X_1, X_2, …X_n. The mean of series is denoted by (X bar); the number of observations by n and the sum of observations by ∑ (sigma) which means sum-up or add-up of all the results.

Formula for calculation of mean is:

$$\text{Mean} = \frac{\text{Total or sum of the observation}}{\text{Number of observations}}$$

The mean is calculated by different methods in two types of series, ungrouped and grouped.

Ungrouped series

In such series the number of observations is small and there are two methods for calculating the mean. The choice depends upon the size of observations in the series.

A. When the observations are small in size, simply add them up and divide by the number of observations.

Example

Tuberculin test reaction of 10 boys is arranged in ascending order being measured in millimeters. Find the mean size of reaction.

3,5,7,7,8,8,9,10,11,12

Mean = $\frac{3+5+7+7+8+9+10+11+12}{10}$ = $\frac{80}{}$ = 8 mm

B. When individual observations are large in size, either use a calculator for addition orproceed as in (A) above. Or assume an arbitrary mean measurement and the assumed mean or working origin (W). Find out the difference between each measurement and the assumed mean or working origin, sum up the difference and divide the total by 'n', i.e., the number of differences or deviations denoted by (small x bar).

Here, X stand for the original series of observations and x for the differences between X and the working origin or assumed mean. Mean of the original observations or X = w + x

Examples

1) Find the mean incubation period of 9 polio cases given below. In this case 20 may be taken as the working origin or working mean and 0 may be put opposite that.

Table 2.3:

X	X-w	X
23	23-20	+3
22	22-20	+2
20	20-20	0
24	24-20	+4
16	16-20	-4
17	17-20	-3
18	18-20	-2
19	19-20	-1
21	21-20	+1
180	+10-10=0	

a. By direct method

Number of observation n = 9

b. By assumed mean (w) method

$$\frac{\sum X}{n} = \frac{180}{9} = 20$$

2. Heights in centimeters for seven school children are given below. Find the mean. 140 may be taken as the working origin to calculate the mean height.

X	X-w	X
148	148-140	8
143	143-140	3
160	160-140	20
152	152-140	12
157	157-140	17
150	150-140	10
155	155-140	15
1065		85

a. By direct method

Number of observations n = 7

$$= \frac{\sum X}{n} = \frac{1065}{7} = 152.1$$

b. By assumed mean (w) method

$$\frac{\sum X - w}{n} = \frac{85}{7} = 12.1$$

= w +(x bar) = 12.1 + 140 = 152.1

If w = 150, calculations will still be easier.

3. The heights of 7 boys aged 14 years are given in cm with fractions in the table that follows. Find the mean. Take 0 for any convenient working origin or assumed mean such as 150 in the example below.

Table 2.4

X (Height)	X-w = x	Sum of x
145.8	-4.2	
146.9	-3.1	-8.5
148.	-1.2	
150.0(w)	0	
152.1	+2.1	
153.6	+3.6	+12.7
157.0	+7.0	

3.5 Grouped Series

When the number of observation is large the data are arranged in groups and frequency distribution table is prepared first. In all grouped series only weighted mean has is prearranged first. In all grouped series only weighted mean has to be found and not the ordinary. Make convenient groups as per the characteristic values and prepare the frequency distribution table. Find the value or weight, contributed by each group separately and multiply the mid value of group with its frequency. Total these product values and then divide by the total number of observations in the sample. This mean is called weighted mean or grand mean or mean of means.

Examples

1. The average income of 10 lady doctors is Rs 400 per month and that of 20 male doctors is Rs 600 per month. Calculate the weighted mean or average income of all doctors.

One may consider the average income of all doctors as Rs 500/ - if he adds the averages of Rs 400 and Rs 600 and then divides by 2. It is, however, advisable to compute after taking into account the incomes or weights contributed by each group of doctors as shown below:
Total income of 10 lady doctors =
= mean income of the group x frequency = xf_1 = 400 x 100 = Rs 4000.
Total income of 20 male doctors = xf_2 = 600 x 20 = Rs. 12000
Total income of all the 30 doctors = 4000 + 12000 = Rs 16000 (sum of weights, i.e. $\Sigma \frac{fx}{n} = \frac{16000}{30} = Rs\ 533.3\ and\ not\ Rs\ 500/-$
Weighted mean is computed in the same way when event in a qualitative data is expressed in percentages.

2. Calculate overall fatality rate in smallpox from the age wise fatality rate given below:

3.

Table 2.5:Fatality Rate

Age Group in Years	No. of Smallpox Cases	Fatality Rate Per cent
0 – 1	150	35.33
2 – 4	304	21.38
5 – 9	421	16.86
Above 9	170	14.17

It will be wrong to compute the overall rate by adding rates and dividing by four.
(35.33 + 21.38 + 16.86 + 14.17) ÷ 4 = 21.94
Instead, we will have to calculate the total deaths in each age group from the fatality per cent given above.

No. of deaths = $\frac{35.33}{100}$ x 150 + $\frac{21.38}{100}$ x 304 + $\frac{16.86}{100}$ x 421 + $\frac{14.17}{100}$ x 170 = 213.06

213.06 deaths have occurred out of 1045 smallpox cases hence overall fatality rate per cent

= $\frac{213.06}{1045}$ x 100 = 20.39%

Calculation can be simplified by directly multiplying the frequency of smallpox cases with the death rate in each age group and then dividing the total deaths by total smallpox cases. This eliminates division as well as multiplication by 100.

$$\frac{150 \times 35.33 + 304 \times 21.38 + 421 \times 16.86 + 170 \times 14.17}{1045} = 20.39$$

Weighted mean is easy when groups are fewer, size of observation is small and they are not in fractions in a frequency distribution table. No working origin is necessary, one has to simply find the value of each group, i.e., weight, add the group weights and divide by the total number in the sample.

4. Find mean days of confinement after delivery in the following series:

Table 2.6: Confinement Days

Days of confinement X	**No of patients f**	**Total days of each group (weights). f X**
6	5	30
7	4	28
8	4	32
9	3	27
10	2	20
	18	137

Apply the formula for the weighted mean $\frac{\sum fx}{n}$ where X denotes the series of observations, i.e., days of confinement and f is the frequency of patients in each group and n is the total frequency, i.e., the number of patients confined.

Mean days of confinement, therefore $= \frac{\sum fx}{n} = 137 \div 18 = 7.61$

Class interval in the days of confinement is taken as one.

5. Calculate mean when class interval is again one but group characteristics and frequencies are in fractions and large in size as in the weight of large number of boys in the table that follows:

Table 2.7: Weight of Large Number of Boys

Weight of Children in kg X	**Midpoint or Average wt. of Each Group X_g**	**No. of Children in Each Group F**	**Weight Contributed by Each Group f X_g**
60 - < 61	60.5	10	10 x 60.5
61 -	61.5	20	20 x 61.5
62 -	62.5	45	45 x 62.5
63 -	63.5	50	50 x 63.5
64 -	64.5	60	60 x 64.5
65 -	65.5	40	40 x 65.5
66 - < 67	66.5	15	15 x 66.5
Total		240	15310.0

Finding the average weight of each group from the original records may be laborious and they may not be available. The groupwise frequency distribution table is prepared as in example 3. To find the average or mean of each group the mid value of the group 60 - < 61. Now the weight contributed by each group (X_g) is found by multiplying with the frequency (f). Apply the formula as in example 3.

$$\frac{\sum f \text{Xg}}{n} = \frac{15310.0}{240} = 63.79$$

This method of calculation is still laborious and can be further simplified by assuming an arbitrary mean or working origin as was done in the ungrouped series when size of observations was large. This working origin will be the mid value of any group which is taken as 0.

The group values above and below this assumed mean (0) may be reduced to working units of 1,2,3, …etc., which are multiples of class or group interval in the frequency distribution table.

Table 2. 8: Frequency Distribution

Weight in kg	Mid value	Frequency or No.	Working units	Group weight	Sum of fx ∑(fx)
X	x_g	f	x	Fx	∑(fx)
60 – 61	60.5	10	-2	10 x -2	
61 – 62	61.5	20	-1	20 x -1	
62 – 63	62.5(w)	45	0	45 x 0	

		75			− 40
63 – 64	63.5	50	+1	50 x 1	
64 – 65	64.5	60	+2	60 x 2	
65 – 66	65.5	40	+3	40 x 3	
66 – 67	66.5	15	+4	15 x 4	
		165			+350
Total		240			+310

Mean ion working units, $x = \frac{310}{240} = 1.29$
Mean in real units, X = w + x = 62.50 + 1.29 = 63.79
Thus, mean in real units = central value (62.5) in real units of the group against which 0 is placed, + X × class interval (1.29 × 1).

Measures of Location – Percentiles

Percentiles Averages discussed so far are measures of central value; therefore, they locate the centre or mid-point of a distribution. It may also be of interest to locate the centre or locate other points in the range. Percentiles do that. They are values of a variable such as height, weight, age, etc., which divide the total observations by an imaginary line into two parts expressed in percentages such as 10 per cent and 90per cent or 25per cent and 75per cent, etc. In all, there are 99 percentiles. Centiles or percentiles are values in a series of observations arranged in ascending order of magnitude which divide the distribution into 100 equal parts. Thus, the median is 50^{th} centile. The 50^{th} percentile will have 50per cent observations on either side. Accordingly, 10^{th} percentile should have 10per cent observations to the left and 90per cent to the right. But for population in India it is not so. If children at age 3.5 years form 10^{th} percentile, it means 10per cent of entire population is below 3.5 years of age and 90per cent is above that age. Age 20 may be median or 50^{th} percentile and divide the people into one half below 20 years and the other half above 20 years of age. In developed countries 10^{th} percentile may be 7.5 years, therefore 10per cent population will be below 7.5 years while in India 7.5per cent years may be 20^{th} or 25^{th} percentile and cover 20per cent or 25per cent of the total population. Thus, percentiles are used to divide a distribution into convenient groups. Those in common use are described below:

Quartiles: They are three different points located on the entire range of a variable such as height – Q_1, Q_2, and Q_3. Q1or lower quartile will have 25per cent observations of heights falling on its left and 75per cent on its right; Q2 or median will have 50per cent observations on both side and Q3 or upper will have 75per cent observations on its left and 25per cent on its right.

Quintiles: Quintiles, four in number divide the distribution into 5 equal parts. So 20^{th} percentile or first quintile will have 20per cent observations falling to its left and 80per cent to its right.

Deciles: Nine in number divide the distribution into 10 equal parts; first decile or 10^{th} percentile will divide the distribution into 10per cent and 90per cent while 9^{th} decile will divide into 90per cent and 10per cent and 5^{th} decile will be same as median. So median of a variable can also be called as second quartile Q_2. 5^{th} decile D_5 or 50^{th} percentile P_{50}.

Arithmetical Method of Finding the Value of any Percentile from the Cumulative Frequency Table

The median, quartile and percentile values of the characteristic can be calculated from the cumulative frequency table. Find the variable group in which the particular observation lies and then raise the lower value of the variable of that group proportionately to the value of that particular observation, with the presumption that the rise in the variable group from the higher value is uniform.

For illustration, median, Q_1, Q_3 and 10^{th} percentile is calculated from the cumulative Frequency.

a. Median (Q_2), i.e.,200/2 or the 100^{th} observation lies in the height group 170 – 172 cm. The cumulative frequency rises by 19 from 81 to 100, i.e., median, or middle observation. For 26 observations, the attribute value as per the table, rises by class interval of 2 cm from 170 cm to 172 cm. Therefore, the proportionate rise in the attribute for 19 observations

$$= \frac{2 \times (100 - 81)}{26} = \frac{2 \times 19}{26} = 1.46$$

Thus, median or second quartile Q_2 value = 170 + 1.46 = 171.46 cm which is almost equal to the graphic value 171.50 (fig. 2.1)

b. First quartile (Q_1), i.e.,200/4 = 50^{th} observation, lies in the group 166 – 168 cm. Cumulative frequency up to height 166 cm is 42.

$Q_1 = 166 + \frac{2 \times (50 - 42)}{19}$ (19 is the group frequency)

$= 166 + \frac{16}{19} = 166.84$ cm

c. Third quartile (Q_3), i.e., 200 × ¾ = 150^{th} observation lies in group 174-176 cm. Cumulative frequency up to height 174 is 136.

$Q_3 = 174 + \frac{2 \times (150 - 136)}{30}$ (30 is the group frequency)

$= 174 + \frac{28}{30} = 174.93$ cm

d. Tenth (10^{th}) percentile, i.e., 200/10 or 20^{th} observation, lies in the group 162 – 164 cm.

$P_{40} = 162 + \frac{2 \times (20 - 10)}{15}$

$= 162 + \frac{20}{15}$

=163.33 cm.

To find how a particular observation or percentile will divide the entire frequency into two parts of percentages, one has to find in which group of the frequency distribution table this particular observation lies and then raise the frequency vale from the lower limit proportionately, again making the presumption that rise is uniform from lower to higher value.

For illustration, let us find how the height 163.33 cm divides the entire frequency. As seen in the example above the height lies in cumulative frequency of group 162 cm to 164 cm, i.e., somewhere in between 10 – 25, proportional rise of frequency from 10 in this group would be calculated as follows:

Rise in frequency for 2 cm height = 15
Rise in height = 163.33 – 162.00 = 1.33 cm
Rise in frequency for the height 1.33 cm
$= \frac{15 \times 1.33}{2} = 10$
Therefore, the number of observations up to the height 163.33 cm = 10 + 10 = 20, which forms the dividing point where 19 observations lie below 20 and 180 above that. Thus, we make use of the cumulative frequency distribution table and not the cumulative frequency polygon.

3.6 Application and Uses of Percentiles

1. Location of a percentile that divides the frequency table into two parts.

   Example
   The 25^{th} percentile is located in the entire height as 166.34 cm graphically and arithmetically. Conversely to find what percentile an observation is

2. Preparation of a standard percentile such as quartile Q_1, median Q_2, etc., for particular age(s), sexes, etc.

Example
Standard median is made use of in the preparation of a weight card, recommended by the Nutrition Subcommittee Indian Academy of Paediatricians to assess the nutritional status of children in the age group 0 – 6 years. The same is used in integrated child development service scheme (India). Median weights of well-to-do Indian children are found at all months from 1 month to 72 months. By joining the median weights plotted month wise on the graph paper, 50^{th} percentile for 72 months is drawn and is used as a standard median. A comparison between well - to – do Indian and American children of corresponding ages revealed that former are as tall and heavy as the latter, boys up to the age 14 years and girls up to the age of 12 years.

Other lines are drawn 80per cent, 70per cent, 60per cent and 50per cent of the standard median weights at different ages. In the diagram thus prepared, the vertical axis indicates the weight from 0 to 22 kg ad horizontal axis gives the age in months from 1 to 72. Weight at any age is plotted to note between which line it lies.

Weight for age falling above the first, i.e., 80per cent line considered normal and that between first line (80% wt.) and second line (70% wt.) is considered in malnutrition Grade 1. Weight falling between 70per cent and 60per cent lines is in Grade 11: between 60per cent and 50per cent is in Grade 111; and below 50per cent is considered in Grade IV. Children falling in nutrition Grades 111 and 1Vare considered as severely malnourished.

3. Comparison of one percentile value of a variable of one sample with that of another sample, drawn from the same population or from different population, median, quartile or any other percentiles can be compared.

Example
Cumulative frequency polygons or ogives for one-year-old Indian and Harvard children are drawn taking an identical sample in both the cases. It is seen that Harvard 50^{th} percentile corresponds to Indian 60^{th} percentile. Weight up to 9 kg covers 50per cent of Harvard children but it covers 60per cent of the Indian children are heavier than the latter.

Thus, size of any percentile of one sample can be compared with other, and deviation found by the difference between the two will be clear at a glance.

4. To study growth in children

Examples

1. A cumulative frequency table of heights of male children who have completed 4th, 5th and 6th year of age, is given below:

Table 2.9: Cumulative Frequency

Age in years	**Height in cm up to different percentiles**						
	5	10	25	50	75	90	95
4	94.0	96.0	99.8	103.7	106.9	110.5	112.9
5	97.3	99.2	102.6	106.2	110.3	114.2	115.8
6	102.8	104.9	108.1	110.4	115.2	11.8.4	119.4

Percentiles at different ages indicate growth which is almost uniform when different percentiles are compared. It is about three cm from age four to five and about four cm from age five to six. The table also shows how different percentiles divide the distribution of children in two parts.

At age four, the height94 cm shows that five per cent children were shorter and 95per cent were taller than height 94 cm. Only five per cent children are taller than 112.9 cm in this group.

2. The 10th percentile height of 200 children at age 4 was 96 cm. Half of the children (100) were put on vitamins A and D and the other 100 children were kept as control and gives placebo.

   After one year, it was found that 10th percentile height as 97.3 cm in control group and 99 cm in experiment group. Thus it was observed that vitamins A and D boosted the height in experiment group in the figure below.

5. As a measure of dispersion: Interquartile and semi-interquartile ranges are sometimes used as measures of dispersion.

4.0 CONCLUSION

Biostatistics is the term used when tools of statistics are applied to the data that is derived from biological sciences such as medicine. Any science demands precision for its development, and so does medical science. For precision, facts, observations or measurements have to be expressed in figures. Everything in medicine be it research, diagnosis or treatment, depends on counting or measurement. High or low blood pressure has no meaning, unless it is expressed in figures. Incidence of tuberculosis or death rate in typhoid is stated in figures. Enlargement of spleen is measured in fingers' breath. Thus medical statistics or biostatistics can be called **quantitative medicine.** Biostatistics may also be called a **science of variation**. The data after collection, lying in a haphazard mass are of no use, unless they are properly sorted, presented, compared, analysed and interpreted. They mean something more than figures, give a dimension to the problem and even suggest the solution. For such a study of figures, one has to apply certain mathematical techniques called **statistical methods**, such as calculation of standard deviation, standard error and preparation of a life table.

A medical student should not depend on a statistician for the statistical analysis. For professional interpretation of his results, he should learn the application of methods himself which do not require knowledge of mathematics higher than what he or she had acquired the application of methods himself which do not require knowledge of mathematics higher than what he or she had acquired at school. Medical Statistics go under different names when applied in different fields such as: Health statistics in public health or community health. Medical statistics in medicine related to the study of defect, injury, disease, efficacy of drug, serum and line of treatment, etc. vital statistics in demography pertaining to vital events of births, marriages and deaths. These terms are overlapping and not exclusive to each other.

5.0 SUMMARY

In this unit, you have learnt the following;

- Application and Uses of Biostatistics as a Science
- Application and Uses of Biostatistics as figures
- Common Statistical terms
- Measures of location-Averages and Percentiles
- Grouped Series
- Application and uses of Percentiles

6.0 TUTOR-MARKED ASSIGNMENT

1) Exhaustively discuss the application and uses of biostatistics as both in science and in figures.
2) A cumulative frequency table of heights of male children who have completed 4th, 5th and 6th year of age , is given below;

Table 2.10: Cumulative Frequency

Age in years	Height in cm up to different percentiles						
	5	10	25	50	75	90	95
4	94.0	96.0	99.8	103.7	106.9	110.5	112.9
5	97.3	99.2	102.6	106.2	110.3	114.2	115.8
6	102.8	104.9	108.1	110.4	115.2	118.4	119.4

a) Attempt to draw the Ogives of the information provided
b) Comment on the percentiles presented at different ages.

7.0 REFERENCES/FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim&Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7thed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

MODULE 5

UNIT 1 WRITING REPORT

CONTENTS

I.0 INTRODUCTION

In this unit, ahead of considering anything about the actual process of getting words onto paper, we look at the construction-“the shape”- of a good report. There are two considerations here:

What makes it work for the reader?
What assist you to compile it quickly and easily?
Of these, the first is the most important, but the factors involved luckily act positively in both cases. The starting point to thinking is to set clear objectives and create very sound structure as we shall briefly discuss in the accompanying paragraphs.

2.0 OBJECTIVES

By the end of this unit, you will be able to:

- state clear objectives
- explain what should constitute the middle portion of the report
- discuss logical structure
- discuss simple format of a good report.

3.0 MAIN CONTENT

3.1 Setting Clear Objectives

The most important thing to settle initially is simply *why* the report is being written. What is it for? Few reports are just "about" something. They may, of course, have various intentions- to inform, motivate and so on, as mentioned in the last chapter- but what matters most is the overall objectives. And this in turn means you must be clear what you want the end results to be after the report is delivered and read.

For example, it is unlikely to be clear objectives to write something "about the possibility of the office relocating". It may be valid to write something to explain why this may be necessary, compare the relative merits of different solution and recommend the best option. Even that may need more specifics within it, spelling out the advantage/ disadvantage to different groups: staffs, customer's etc. who may each be affected in different ways.

Objectives should be defined from *standing point of readers.*
You need to consider:

- Which particular people the report is for?
- Whether the group is homogenous or if multiple needs must be met?
- The reason these people needs the report?
- What they want in it, and in what detail?
- What they do *not* want;

The result they look for (what they want to understand, what action they want to take, or what decision they will be able to make).

It follows that it may well help to know something about the recipients of any reports that write. You may, of course, know them well; for example, they may be colleagues that you work with closely. If not, ask yourself:

- What kind of people are they (male/female, young/old)
- How well do you know them?
- What is their experience of the report topic?
- What is their likely attitude to it? (e.g. welcoming/hostile)
- What is their personal involvement? (How does the issue affect them?)
- How do they rank the importance of the topic?
- Are they likely to find the topic interesting?
- Are they likely to act as a result of reading it?

Everything that follows, what you write, how you write it and how you arrange it, is dependent on this first premise –a clear objective is literally the foundation upon which a good report is based. We will return to this, and to exactly how you set such an objective, in considering preparation for writing in this next chapter. Meanwhile we turn to the actual shape of the report itself.

3.2 A Sound Structure

1) The simplest structure one can imagine is a beginning, middle and the end. Indeed, this is what a report must consist of, arguments or case it presents may be somewhat more complex. This falls naturally into four parts:
   - Setting out the situation.
   - Describing the implication.
   - Reviewing the possibilities
   - Making a recommendation.
   -

2) The situation: this might refer to both the quantity and importance of written communication around, and outside, the organisation. Also to the fact that writing skills were poor, and no standard were in operation, not had any training ever been done to develop skills or link them to recognised models that will be acceptable around the organisation.

The implication: these might range from a loss of production (because document took too long to create and had to constantly be referred back for clarification), to inefficiencies or worse resulting from misunderstood communications. It could also include dilution or damage to image because of poor document circulating around the organisation, perhaps to customers.

The possibilities: here, as with the argument, there might be many possible courses of action, all their own mix of pros and cons. To continue the example, it might range from limiting report writing to a small core of group of people, to reduce paper work completely or setting up a training programme and subsequent monitoring system to ensure some improvement took place.

The recommendation: here the “the best” option needs to be set out. Or, in some reports, a number of options must be reviewed from which other can choose. Recommendations needs to be specific: addressing exactly what should be done, by whom, when, alongside such details as cost and logistics.

At all stages generalisations should be avoided. Reports should contain facts, evidence, and sufficient "chapter and verse" for those in receipt of them to see them as an appropriate basis for decision or action.

With the overall shape of the argument clearly in mind we can look in more details at the shape of the report itself. The way in which it flows through from the beginning to the end is intended to carry the argument, make it easy to follow and to read, and to make it interesting too, a necessary, along the way.

The three parts fit, unsurprisingly, the old and useful maxim about the communications usually abbreviated to: "tell" em, tell" em and tell "em". In full this says: tell people what you are going to tell them- *the introduction*, tell them in details- *the body of the report*, and then tell them what you have told them- or *summarise.*

3.3 First, Essential Features of the Beginning

This must start by addressing the stance of the readers. What will they be thinking as they start reading? -*will it be interesting, Readable? Will it help me? Is it important?* -And will it distract them from anything else going on around them, engaging their concentration so that they give it their attention?

They have their agenda, wanting the report to be succinct etc. as mentioned earlier; essentially they will only give it real consideration if they find it understandable, interesting and a good fit with their situation. They do not want to find it inappropriate. It should not: confuse them, blind them with science/technicalities or jargon lose them in an impenetrable structure (or lack of it), or talk down to them.

Judgments are made very quickly. In the first few lines a view is adopted that colour their reading of the rest of the document. First impression last, as the old saying has it, so this stage is very important and may need disproportionate thought to get it phrased and constructed just right.

The beginning must act as an introduction, which must:

- Set the scene (this can include linking to terms of reference or past discussion that can promote the report to be written);

- State the topic and theme (and maybe treatment);
- Make the objectives clear;
- Begin to get into the topic, creating a thread that draws the reader through the first part to the core of the report;
- Position it as appropriate for the readers (who must feel they are, as it were, eavesdropping on something meant for others).
- At the same time, the beginning will inevitably say something about the writer, and therefore needs to reflect anything you want readers to feel (that you are expert, professional or whatever) and not put out any untoward message (too much jargon may say "this person does not understand the needs of their readers"). So this element must be injected, something we return to.

If it is to earn a reading a report must get quickly to the point. This does not preclude setting the scene. A report might start: "this report sets out to demonstrate how the organisation can cut cost by 10 per cent, without sacrificing quality". After this, and perhaps a little more, had got readers wanting to know how, it may be necessary to go back and set the scene in terms that reflects an analysis of current expenditure. But people know where the report is going- they will go through the text more easily once a desirable intention has been spelt out.

The tone of a report also needs to show itself at this first stage. Just as presenters needs to establish a rapport with their audience, so a report receives continuing attention if it comes over as necessary, useful, written for a purpose, written with conviction, written by someone the reader want to listen to, and- above all- written with understanding of, and concern for, its readers.

Get off to a good start and any continuing task is then often easier. This certainly applies to writing. Feeling you have got a good beginning breeds confidence in what must follow. And so too with reading: if a document starts well, people read on wanting to rest to match the early acceptability. There are plenty of frustration in corporate life, something that looks set to make life a little easier is very soon recognised and appreciated.

One of the reasons that what is often called the executive summary (a summary that is placed at the beginning rather at the end) often works well is that it meets many of the criteria for the beginning stated. It interests the reader who then reads on to discover the detail and see how and why the stated conclusions have been arrived at.

3.4 The Middle

This is the core of the document. It is where the greatest amount of the content is to be found, and hence it has the greatest need for structure and organisation. The key aims here are to:

- Put over the detail of the report's message;
- Maintain, indeed develop, interest;
- Ensure clarity and a manner appropriate to the reader.

It may be necessary to go further. It is here that the report may seek acceptance and, conversely, set out to counter people disagreeing with or rejecting what it has to say. At the same time any complexity must be kept manageable. Doing this, necessitates the simple practice what I preach, are a number of points all aimed at keeping this core section on track.

3.5 Putting Over the Content

A logical structure: selecting, and describing to your readers, a way through the content (for example, describing something in chronological order).

"Signposting" intentions; knowing what is coming (and why) makes reading easier. This is why many documents need a content page, but it can also be done within the text- "we will review the project in terms of three keys factors: timing, cost and staffing. First, timing……' (Perhaps this is followed by the heading 'timing'). This is something that is difficult to overdo, the clarity it promotes and the feeling of having what is being read in context of what is to come is appreciated.

Using heading (and sub heading); this is not only effectively a form of signposting, it breaks the text visually and makes it easier to work through a page (contrast the style of a modern business book, such as this, with the kind of dense textbook many of us suffered with at school).

Appropriate language: this is important at every stage
Using graphics (visual graphic devices): this encompasses two types of factors- such things as bold types, capital letters etc. and illustrations, including graphs, tables, charts etc.

3.6 Format of Research Report

- Technical report:
- Interim report:
- Summary report:
- Steps in writing research report
- The type and scope of report
- Format of research report;
- Acknowledgements,
- Table of contents,
- List of tables, list of graphs and charts,
- Abstract of synopsis
- Dedication page

ii) Body of the report:

- Introduction,
- Theoretical background of the topic,
- Statement of the problem,
- Review of literature,
- The objectives of the study,
- Hypotheses to be tested,
- Definitions of the concepts,
- Models, if any,
- Capitalisation,
- The design of the study,
- Methodology,
- Formulating a research problem,
- Hypothesis,
- Results / Findings,
- Discussion,
- Summary, conclusions and recommendations,
- Terminal items: bibliography, appendix,
- Copies of data collection instruments, technical details on sampling plan,complex tables glossary of new terms used in the report.

4.0 CONCLUSION

A good report has two inherent important questions to brood over and that must be rapidly considered and answered, viz;-:
What makes it work for the reader?

What assist you to compile it quickly and easily? Of these, the first is the most important, but the factors involved luckily act positively in both cases. The starting point to thinking is to set clear objectives and create very sound structure.

5.0 SUMMARY

In this unit we have learnt about the following with respect to creating or generating a good report:

- Ability to set clear objectives
- Creating a sound structure
- Including the essential features of the Beginning
- What should constitute the middle portion of the report?
- Putting over the content, viz;-
  - Logical structure
  - Sign posting
  - Using heading and subheading
  - Using appropriate language
  - Graphics
- Knowing simple format of a good report

6.0 TUTOR-MARKED ASSIGNMENT

1. Clearly show the format of creating a good report.
2. As a newly appointed public health officer in your locality besieged with an outbreak of Lassa fever, write a full account of how you would present your prime report to your health commissioner.
3. Write briefly on the following with respect to the content of a good public health report;
   i) Logical structure
   ii) Sign posting
   iii) Using heading and subheading
   iv) Using appropriate language
   v) Graphics

7.0 REFERENCES/FURTHER READING

Bushit, D.B. & Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim & Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6th ed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5

UNIT 2 PROPOSAL WRITING

CONTENTS

1.0 INTRODUCTION

Proposals written to solve or address public Health problems and challenges are very important documents. Everything we have already discussed in report writing is equally applicable here with much emphasis on the following; opening must command attention, be visibly, and strikingly persuasive, establish interest and lead into the main text, making people want to read on. As the introduction has to undertake a number of important, yet routine, tasks, ahead of them it may be best to start with a sentence (or more) that is interesting, rings bells with the client's needs and wants and set the tone for the document.

The accompanying pattern may be very useful:

- Establish the background of the proposal;
- Refer to past meeting and discussions;
- Recap decisions made to date;
- Quote experience;
- Acknowledge terms of reference;
- List the name of those who were involved in the discussions and/or preparation of the document.
- As none of this is as interesting as what will follow, this section should concentrate on essentials and kept short.
- The final words should act as a bridge to the next section.

2.0 OBJECTIVES

Bythe end of this unit, you will be able to:

- explain statement of need when writing proposals
- discuss proposal content.

3.0 MAIN CONTENT

3.1 Statement of Need When Writing Proposals

This section needs to set out, with total clarity, the brief in terms of the needs of the customer. It describes the scope of the requirement and may well act as to recap and confirm what was agreed on at the prior meeting that the proposal was intended to cover.

It is easy to ask why this should be necessary. Surely the customer knows what he or she wants? Indeed, he or she has perhaps just spent a considerable amount of time telling you exactly that. But this statement is till important.

Its role is to make clear that you do have complete understanding of the situation. It emphasises the identity of views between two parties, and gives credibility to your later suggestions by making clear that they are based firmly on the real needs that exist. Without this it might be possible for the customer to assume that you are suggesting what is best (or perhaps not profitable) for you; or simply making a standard suggestion.

This section is also key if the proposal is to be seen by people who were not party to the original discussions; for them it may be the first clear statement of this picture.

Again, this part should link naturally into the next section.

3.2 Proposal Content

While the form and certainly the content of a proposal can vary, the main divisions are best described as:

- The introduction (often preceded by a content page);
- The statement of need;
- The recommendation (or solution);

- Areas of details (such as costs, timing, logistics, technical specification);
- Closing statement (or summary);
- Additional information (of prime or lesser importance- in form of appendices).

Each may need a number of subheadings and their lengths may vary with context, but they form a convenient way of reviewing the key issues about the construction of a proposal and are thus commented on below. A proposal of any complexity needs the equivalent of a book's title page. This states whom, or which organisation, it is for, what it is about and whom it is form. This page can also give the contact details- address etc.of the proposer (which, if not here, must certainly be in the proposal somewhere) and some proposal writers like to feature the logo of the recipient organisation on it, as well as his or her own.

A front sheet giving the contents, and page number should follow this. It may make it look more interesting if there are subheading as well as main heading have to be bland, e.g. 'the introduction'. Action words- making, creating etc.- should be preferred.

The heading that follows below are descriptive of the functions and roles of the sections, not recommendation for heading you should necessarily use.

3.3 Recommendation

This may well be the longest section needs to be logically arranged and divided (as do all the sections) to make it manageable. Here you state what you feel meets the requirements. This may be: standard, in the sense that it is a list of, for example, recommend equipment and spares which are all items drawn from published information such as a catalogue. Or it may be spoke, as with the approach a consultant might set out to instigate a process of change or implement training.

In either case this section needs to be set out in a way that is 'benefits-led', spelling out the advantages and making clear what the solution will mean to, or do for, the individual customer as well as specifying the technical features.

Remember, the sales job here is threefold: to explain, to do so persuasively and also to differentiate. Never forget, when putting together a proposal, that you may well be in competition and what you present will be compared with the offerings of others.

A focus on the customer's needs is usually the best way to ensure the readers' attention; nothing must be said that does not have clear customer relevance. One further emphasis is particularly important here: individuality. It is very easy to store standard document on disk these days, and it's indeed possible to edit one proposal into a new version that does genuinely suit a similar need elsewhere (but double, double check that you have changed the customer's name!). It must not seem standardised though. This is sufficiently important to reemphasise: it must never seem standard in any sense. Customers may well know that you must get many similar requests, but will still appreciate clear signs that you have prepared something tailored just for them.

Only when this section has been covered thoroughly should you move on to costs. Only when the customer appreciates exactly what values and benefits are being provided can price be considered in context.

3.4 Costs

These must be stated clearly, not look disguised though certain techniques for presenting the figures are useful, e.g. amortising costs.

All the necessary detail must be there, including any items these are:

- Options;
- Extras;
- Associates expenses.

These must be shown and made clear. I know of one company that lost the contract when one of their executives met the managing director of the customer at a railway station and it was clear that all travel- which was agreed- was being billed for first class – which had neither discussed nor specified.

This no place for a treatise on pricing policy, but note that:

- Price should be linked as closely as possible to benefits;
- This section must establish or reinforce that you offer value for money;
- Invoicing details and trading often need including, and must always be clear; mistakes here tend to be expensive. (in the UK remember to make clear whether price is inclusive of VAT);
- Overseas, attention must be given to currency considerations;
- Comparison may need to be made with the competition;

- Range figure (necessary in some businesses) must be used carefully (do not make the gap too wide and never go over the upper range figure).Look carefully at how you arrange this section. It is only realistic to assume that some readers will look at this before reading anything else. Certainly for them need to be sufficient explanation, cost justification and –above all- clear benefits, linked in here. Just the bald figures can be very off putting.

3.5 Details

There are additional topics that may be necessary to deal with here, as mentioned above: timing, logistics, staffing etc. sometimes they are best combined with costs and timing go well together, with perhaps one other separate, numbered, section dealing with any final topics before moving on. The principles here are similar to those for handling costs. Matters such as timing be made completely clear and all possibilities of misunderstanding or omission avoided.

3.6 Closing Statement

The final section must be act to round off the document and it has a number of specific jobs to do. It is first, and perhaps, most important, task to summarise. All the thread must be drawn together and key aspects emphasised. This fulfils a number of purposes:

- It is a useful conclusion for all readers and should ensure the proposal ends on a note that they can easily agree is an effective summary. Because this is often the most difficult part of the document to write, also a part can impress disproportionately. Readers know good summarising is not easy and they respect the writer who achieves it.
- It is useful too in influencing others, around the decision maker, who may study the summary but not go through the whole proposal in detail.
- It ensures the final word, and the final impression left with the reader, is about benefits and value for money.

In addition, it can be useful to:

- Recap key points (as well as key benefits);
- Stress that the proposal are, in effect, the mutual conclusion of both parties (if this is so);
- Link to action, action dates and points and people of contact (though this could equally be dealt with in the covering letter);

- Invoke a sense of urgency (you will normally hope for things to tie down promptly, but ultimately must respect the prospect's timing).

Remember that this summary may have to work in content of the so-called 'executive summary', which is placed at the start of the document to do much the same job.

3.7 Additions

The key thing here is appendices. It is important that proposals, like any document, flow. The argument they present must proceed logically and there must be no distractions from the developing pictures. Periodically, there is sometimes a need to go into deep detail. Especially if this is technical, tedious or if it involves numerous figures – however necessary the content may be – it is better not to let such detail slow and interrupt the flow of the argument. Such information can be usually referred to as the appropriate point, but with a note that the 'chapter and verse' is in an appendix. Be specific, saying for example: *costs and timing,* which appears. This arrangement can be used for variety of elements: terms of reference, contract details, worked examples, graphs and figures, tables and so on. Each of the major sections should be appropriately, and if possible, interestingly titled and you may sensibly start each main section on a new page, certainly with a proposal of any length. Language and layout are also very important**.**

4.0 CONCLUSION

Proposals are very important documents written to address public health and some other medical problems likewise, social, economic and political challenges. Key issues and facts we discussed in report writing are equally applicable in proposal writing with much emphasis on the following; opening section of any particular proposal must command attention, be visibly, and strikingly persuasive, establish interest and lead into the main text, making people want to read on. Always best to start with a sentence (or more) that is interesting, rings a bell with the client's needs and wants meaningfully documented and properly addressed. The patterns shown below are very useful in proposal writing, viz;-:

- Establish the background of the proposal;
- Refer to past meeting and discussions;
- Recap decisions made to date;
- Quote experience;
- Acknowledge terms of reference;

- List the name of those who were involved in the discussions and/or preparation of the document.
- As none of this is as interesting as what will follow, this section should concentrate on essentials and kept short.
- The final words should act as a bridge to the next section**.**

5.0 SUMMARY

In this unit, we have learnt the following with regard to proposal writing in public health.

- Statement of Need When writing Proposals
- Proposal Content
- Recommendation
- Costs
- Details
- Additions

6.0 TUTOR-MARKED ASSIGNMENT

1. What is meant by the term "statement of need in proposal writing"?
2. What differentiate report from proposal writings?
3. You have the burning desire to submit a proposal to a donor agency for fund in order to execute a communal project on malaria disease as the public health officer in charge. State fully your detailed submissions to the agency.
4. Discuss cost implication and its analysis in proposal writing.

7.0 REFERENCES/ FURTHER READING

Bushit, D.B. &Visweswara, Rao. (2009). Basic Principles of Medical Research, Text Bk. of Biostatistics, A manual of Statistical Methods for use in Health, Nutrition and Anthropology. (2nd ed.).Jaypee Brothers Medical Publishers Ltd. New Delhi: ISBN 81-8448-055-5.

Geraint, H. Lewis, Jessica,Sheringham, Kanwal,Kalim& Tim, J.B. Crayford (n.d) Mastering Public Health Royal Society of Medicine Press Ltd 2008 ISBN-13 978-1-85315-781-3

Kothari, C.R., Gauray, Garg (2015). Researcher Methodology (methods and techniques) (3rd ed.). New Age International Publishers ISBN: 978-81-224-3623-5

Murray, R. Spiegel, Larry, J. Stephen, (n.d). Ph.D., Schaum's outline of theory and problem of statistics, the McGraw Hill companies 2008ISBN: 978-0-07-148584-5

Mahajan, B. K. (n.d).Methods in Biostatistics (6thed.).Jaypee Brothers Medical Publishers Ltd B-3 EMCA House, 23/23B Ansari Road, Daryaganj Post Box 7193, New Delhi, India.ISBN: 81-7179-520-X

Patrick, Forsyth (n.d). How to Write Reports and Proposals (7th ed.).Kogan Page Limited 120 Pentonville Road London 2006, ISBN: 10 0 7494 4552 1, ISBN: 13 978 0 7494 455 2

Visweswara, Rao (2009). Text Bk. of Biostatistics, A Manual of Statistical Methods for use in Health, Nutrition and Anthropology.(2nd ed.). Jaypee Brothers Medical Publishers Ltd. New Delhi, India. ISBN 81-8448-055-5