LECTURE NOTES

For Health Science Students

Research Methodology

Ethiopia Public Health
Training Initiative

Getu Degu
Tegbar Yigzaw

University of Gondar

In collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education

2006

Funded under USAID Cooperative Agreement No. 663-A-00-00-0358-00.

Produced in collaboration with the Ethiopia Public Health Training Initiative, The Carter Center, the Ethiopia Ministry of Health, and the Ethiopia Ministry of Education.

PREFACE

This lecture note on research methodology is primarily aimed at health science students. It is also hoped to be useful for other individuals who would like to understand the basic principles and undertake health research. There is a strong belief that it will serve as a guideline for undergraduate health science students as they are required to identify the most important health problems and carry out some research work.

Chapter one deals with the general introduction and it is devoted to giving basic definitions of important terms and characteristics of research in general and health research in particular. Chapter two gives the guidelines useful for the identification and selection of a research topic. The questions relating to whether a research problem is adequately analyzed and whether it is clearly stated are addressed in Chapter three. Chapters four and five deal with literature review and the development of research objectives, respectively. A special emphasis is given to Chapter six which is the Chapter that contains the many elements of the "Methods" section of a research proposal. Chapter seven deals with the development of a Work plan and the preparation of a budget for a given study.

A summary of the major components and outline of the different phases in a research process (proposal development, fieldwork and report writing) is given in Chapter eight. This Chapter presents the format that an investigator may follow when writing the final draft of his/her health research proposal. It also gives the guidelines for writing a report. The last chapter is devoted to giving a brief account of the definitions of common terms applied in computer use and the application of some statistical packages. A special emphasis is given to Epi6.

In general, this lecture note tries to cover the three major components of a research process: development of the research proposal, fieldwork (data collection) and write-up of the scientific report. General learning objectives followed by introductory sections which are specific to each chapter are placed at the beginning of most of the chapters. The lecture note also includes a number of exercises for the students so that they can examine themselves whether they have understood the topic under consideration. It is assumed that this lecture note on research methodology will be given to health science students who have taken basic Epidemiology and Biostatistics courses.

It is also important to note that this lecture note focuses on quantitative research. When the point of discussion refers to qualitative research, it would be clearly shown so as to avoid the confusion that may arise. A few reference materials are given at the end of the lecture note for further reading.

ACKNOWLEDGMENTS

We would like to thank the College of Medicine and Health Sciences (University of Gondar) for allowing us to use the resources of the institution while writing this lecture note. We are highly indebted to the Carter Center (Ethiopian Public Health Training Initiative) without whose support this material would have not been written. In particular, we are very grateful to Ato Aklilu Mulugeta (from the Carter center) for his uninterrupted follow up and encouragement. We would like to extend our gratitude and appreciation to Dr. Getnet Mitikie and Dr. Mesganaw Fantahun of Addis Ababa University Associate and Assistant Professors respectively for their critical reviews and valuable comments on the initial draft of these teaching materials. .

TABLE OF CONTENTS

CHAPTER ONE
INTRODUCTION TO RESEARCH

1.1 Learning Objectives

After completing this chapter, the student should be able to:

1. Define research in general and health systems research in particular
2. Enumerate the characteristics of research
3. Identify the different types of research
4. List the essential features of health systems research
5. Describe the broad divisions (steps) involved in the research process
6. Explain the roles of research in development

1.2 Introduction

The ultimate goal of any national health-development process is to enable its people to reach a level of health that enables them to make meaningful participation in the social and economic life of the community in which they live. To attain this objective, countries should decide on the best approaches to adopt. However, this requires detailed and accurate information on the existing health systems of these countries. Unfortunately, such information is often lacking , inadequate, or unreliable. As a result, decisions are based on assumptions and unjustified conclusions and often result in inappropriate policy choices. In this regard, the search for scientific knowledge and information should be strongly supported.

Research in the context of public health thus aims to provide all aspects of information necessary for planning and the effective implementation of a health system. For all communities, whether affluent or poor, health research is the top priority. The research questions are formidable: how to join with policy makers and communities in assessing priority needs, planning, financing and implementing programs, and evaluating them in terms of coverage, efficiency and effectiveness.

1.3 Definition and characteristics of research

Definition: Research is a scientific inquiry aimed at learning new facts, testing ideas, etc. It is the systematic collection, analysis and interpretation of data to generate new knowledge and answer a certain question or solve a problem.

Characteristics of research

- It demands a clear statement of the problem
- It requires a plan (it is not aimlessly " looking" for something in the hope that you will come across a solution)
- It builds on existing data, using both positive and negative findings
- New data should be collected as required and be organized in such a way that they answer the research question(s)

1.4 Types of research

Research is a systematic search for information and new knowledge. It covers topics in every field of science and perceptions of its scope and activities are unlimited. The classical broad divisions of research are: basic and applied research. The **basic research** is necessary **to generate** new knowledge and technologies to deal with major unresolved health problems. On the other hand, **applied research** is necessary **to identify** priority problems and to design and evaluate policies and programs that will deliver the greatest health benefit, making optimal use of available resources.

Quantitative and Qualitative researches: Early forms of research originated in the natural sciences such as biology, chemistry, physics, geology etc. and was concerned with investigating things which we could observe and measure in some way. Such observations and measurements can be made objectively and repeated by other researchers. This process is referred to as "quantitative" research.

Much later, along came researchers working in the social sciences: psychology, sociology, anthropology etc. They were interested in studying human behaviour and the social world inhabited by human beings. They found increasing difficulty in trying to explain human behaviour in simply measurable terms. Measurements tell us how often or how many people

behave in a certain way but they do not adequately answer the “why” and “how” questions. Research which attempts to increase our understanding of why things are the way they are in our social world and why people act the ways they do is “qualitative” research.

Qualitative research is concerned with developing explanations of social phenomena. That is to say, it aims to help us to understand the world in which we live and why things are the way they are. It is concerned with the social aspects of our world and seeks to answer questions about:

- Why people behave the way they do
- How opinions and attitudes are formed
- How people are affected by the events that go on around them
- How and why cultures have developed in the way they have

Qualitative research is concerned with finding the answers to questions which begin with: why? How? In what way? Quantitative research, on the other hand, is more concerned with questions about: how much? How many? How often? To what extent? etc.

Public health problems are complex, not only because of their multicausality but also as a result of new and emerging domestic and international health problems. Social, economic, political, ethnic, environmental, and genetic factors all are associated with today’s public health concerns. Consequently, public health practitioners and researchers recognize the need for multiple approaches to understanding problems and developing effective interventions that address contemporary public health issues. Qualitative methods fill a gap in the public health toolbox; they help us understand behaviors, attitudes, perceptions, and culture in a way that quantitative methods alone cannot. For all these reasons, qualitative methods are getting renewed attention and gaining new respect in public health.

A thorough description of qualitative research is beyond the scope of this lecture note. Students interested to know more about qualitative methods could consult other books which are primarily written for that purpose. The main purpose of this lecture note is to give a detailed account on the principles of quantitative research.

Health research

Health research is the application of principles of research on health. It is the generation of new **knowledge** using scientific method to identify and deal with health problems. Knowledge, both generalizable worldwide and locally specific, is essential to effective action for health. Worldwide knowledge is the basis on which new tools, strategies, and approaches are devised that are applicable to health problems facing many countries. Local knowledge, specific to the particular circumstances of each country can inform decision regarding which health problems are important, what measures should be applied and how to obtain the greatest health benefit from existing tools and limited resources. In this regard, health research is both global and local in nature.

In most cases, health research has been divided into three overlapping groups.

Essential health research: Consists of activities to define the health problems of a given country or community, to measure their importance and to assure the quality of activities to deal with them. Much of this research comes within the category of health service research but there will be elements of clinical research and development of technology, depending on the situation. The information, which may be obtained in a number of ways, is essential and specific to each country for planning and monitoring health services. Some of the research conclusions, however, may be generalized and applicable to other areas.

Clinical research: In its widest sense, this group of topics ranges from studies of the prevention and diagnosis of diseases through new methods of treatment to problems of care and rehabilitation. The sophistication will vary from problem to problem and there will be overlap with the fields of essential and biomedical research. Some of the research will be mainly of local importance; much will be useful for other individuals in other countries. Examples include clinical trials of disease prevention and the design of new chemotherapeutic agents. Wherever clinical facilities exist, there is a potential for clinical research.

Biomedical research: It is the most basic part of health research which demands more resources, facilities and skilled investigators. The results of biomedical research are more often of universal importance and thus of general significance.

During the past two decades, concepts and research approaches to support health development have evolved rapidly. Many of these have been described by specific terms such as operations research, health services research, health manpower research, policy and economic analysis and decision-linked research. Each of these has made crucial contributions to the development of health research.

1.5 Health systems research

It is a component of health research. Research that supports health development has come to be known as Health Systems Research. It is ultimately concerned with improving the health of a community, by enhancing the efficiency and effectiveness of the health system as an integral part of the overall process of socioeconomic development.

Definition of "health system"

A health system may be described as:

- A set of cultural beliefs about health and illness that forms the basis for health-seeking and health-promoting behaviour.
- The institutional arrangements within which that behaviour occurs; and
- The socioeconomic (political) physical context for those beliefs and institutions.

In short, it consists of what people believe and know about health and illness and what they do to remain healthy and cure diseases. Beliefs and action are usually closely connected. For example, if in a society people perceive germs as the cause of disease, they will look for modern (biomedical) health care.

The institutional arrangements within which the health-seeking and health-promoting behaviour occurs may include:

1. **The individual, family and the community**
2. **Health care services** ——— private sector: traditional and modern medical practice (legal or illegal)

   Public (governmental) sector → Health workers, health institutions, etc.
3. **Health related sectors** ⟶ education, agriculture, etc.
4. **The international sector,** including bilateral and multilateral donor agencies (UNICEF, WHO, etc.) that may support health as well as **Essential Features of Health Systems Research (HSR)**

Bearing in mind that HSR is undertaken primarily to provide information to support decision-making at all levels that can improve the functioning of the health system, some of the essential features are summarized as follows:

- HSR should focus on **priority problems**.
- It should be action oriented (i.e., aimed at developing solutions)
- An integrated **multidisciplinary** approach is required (research approaches from many disciplines)
- The research should be **participatory** in nature (from policy makers to community members)
- Research must be **timely.**
- Emphasis should be placed on comparatively **simple, short-term research designs** that are likely to yield practical results.

1. The principle of **cost-effectiveness** is important in the selection of research projects.

- Results should be presented in **formats most useful for administrators, decision-makers and the community.**
  - A clear presentation of results with a summary of the major findings adapted to the interests of the party being targeted by the report.

- Honest discussion of practical or methodological problems that could have affected the findings.
- Alternative courses of action that could follow from the results and the advantages and drawbacks of each.

9. **Evaluation of the research undertaken** - An HSR project should not stop at finding answers to the research questions posed, but include an assessment of what decisions have been made based on the results of the study. This is the ability of research findings to influence policy, improve services and contribution to the betterment of health.

1.6 Main components of any research work

I. Preparing a research proposal

II. Fieldwork (i.e., data collection)

III. Analyzing data and preparing a research report

N.B.

The roles of health managers and the community should be identified in the various phases of the research process.

1.7 Exercises

1. The health of any community depends on the interaction and balance between the health needs of the community, the health resources that are available, and the selection and application of health and health related interventions. Discuss!
2. To invest in research is to invest for a better future. Does this statement sound true? Justify your answer.
3. Describe the characteristics of HSR by giving your own examples.

CHAPTER TWO

TOPIC SELECTION

2.1 Learning objectives

After completing this chapter, the student should be able to:

1. Examine the cyclical nature of the development of a research proposal
2. Describe the principles underlying whether a problem situation is researchable.
3. List the criteria for selecting a research topic.
4. Identify and select his/her own topic (health problem) for research based on certain guidelines.

2.2 Introduction

The development of a health project goes through a number of stages. Formulation of the research proposal is the major task in the process of developing a research project. The proposal draws on all the preparatory steps of the research process and pulls them together in a document describing the rationale and the methodology proposed for research. The proposal is a basis for approval and funding. After approval, the proposal is used as a blueprint during implementation of the project. It should be noted that development of a research proposal is often a cyclical process. The process is not always linear. It is a usual practice to go up and down on the developed proposal and make the necessary revisions.

Is there evidence to indicate that the research proposal focuses on a problem of priority importance? Was the given health problem identified by relevant groups of the health system? Was the problem adequately analysed to include all possible contributory factors from different sectors? Was it clearly stated? These questions should be clearly answered before trying to develop the research proposal. The sections that follow are devoted to giving the guidelines useful for identification, selection, analysis and statement of the given problem.

2.3 Problem identification

If the answer to the research question is obvious, we are dealing with a management problem that may be solved without further research. A number of research questions could be presented that may be posed at the various levels of the health system.

Whether a problem requires research depends on *three* conditions:

I) There should be a perceived difference or ***discrepancy between what it is and what it should be;***

II) The reason(s) for this difference should be ***unclear*** (so that it makes sense to develop a research question); and

III) There should be more than one possible and plausible answer to the question (or solution to the problem).

example1:

Problem situation: In district **" Y "** a report showed that in the first month there were 500 children under one year old who started immunization, but at the end of the year it was found out that there were only 25 children who completed their vaccination.

Discrepancy: All the 500 children at district **"Y "**should have completed their vaccination but only 5% out of those who started vaccination have completed.

Problem (research) question: why only 5% of the children completed their vaccination?

Definite answer: Out of the 1 hospital, 2 health centers and 10 health stations found in district **"Y"** only 2 health stations were functioning, the rest were closed due to insecurity in the area.

In the above example, assuming that all the given facts are true, there is no need of undertaking a research, since definite answer is obtained to the problem situation.

Example 2:

Problem situation: In district "Z" (population 150,000) there are 2 health centers, 1 hospital and 15 health stations and all of them function smoothly. However, at the end of the year it was found that the EPI coverage was only 25%.

Discrepancy: Although district **"Z"** had 100% availability of health services and at least 80% of the children should have had full vaccinations the EPI coverage was only 25% as seen above.

Problem question: What factors influence the low EPI coverage in district **"Z"?**

Possible answers:

- Mothers might have problems for not attending in the EPI sessions.
- The **MCH, EPI, OPD, CDD**, etc… programmes might not have been integrated; hence children might have missed opportunities in getting immunization.
- The follow up of defaulting children might not be effective and other reasons.

Thus, the above problem situation is researchable.

2.4 Criteria for prioritizing problems for research

Each problem that is proposed for research has to be judged according to certain guidelines or criteria. There may be **several ideas to choose** from.

Before deciding on a research topic, each proposed topic must be compared with all other options.

The selection and analysis of the problem for research should involve those who are responsible for the health status of the community. This would include managers in the health services, health-care workers, and community leaders, as well as researchers.

The guidelines or criteria given below can help in the process of selection.

a) Criteria for selecting a research topic

1. **Relevance:** The topic you choose should be a priority problem:

   Questions to be asked include:

   - ***How large or widespread is the problem?***
   - ***Who is affected?***
   - ***How severe is the problem?***

2. ***Avoidance of duplication***: Investigate whether the topic has been researched.

   If the topic has been researched, the results should be reviewed to explore whether major questions that deserve further investigation remain unanswered. If not, another topic should be chosen.

3. **Feasibility:** Consider the complexity of the problem and the resources you will require to carry out the study.

   Thought should be given first to personnel, time, equipment and money that are locally available. In situations where the local resources necessary to carry out the project are not sufficient, you might consider sources available at the national level.

4. **Political acceptability:** It is advisable to research a topic that has the interest and support of the authorities. This will facilitate the smooth conduct of the research and increases the chance that the results of the study will be implemented.

5. **Applicability of possible results and recommendations**

   Is it likely that the recommendations from the study will be applied? This will depend not only on the blessing of the authorities but also on the availability of resources for implementing the recommendations**.**

6. **Urgency of data needed**

How urgently are the results needed for making a decision? Which research should be done first and which can be done late?

7. **Ethical acceptability**

We should always consider the possibility that we may inflict harm on others while carrying out research. Therefore, it will be useful to review the proposed study.

b) Scales for rating research topics

Relevance

1 = Not relevant

2 = Relevant

3 = very relevant

Avoidance of duplication

1 = Sufficient information already available

2 = Some information available but major issues not covered

3 = No sound information available on which to base problem-solving

Feasibility

1 = Study not feasible considering available resources

2 = Study feasible considering available resources

3 = Study very feasible considering available resources

Political acceptability

1 = Topic not acceptable

2 = Topic somewhat acceptable

3 = Topic fully acceptable

Applicability

1 = No chance of recommendations being implemented

2 = Some chance of recommendations being implemented

3 = Good chance of recommendations being implemented

Urgency

1 = Information not urgently needed

2 = Information could be used but a delay of some months would be acceptable

3 = Data very urgently needed for decision-making

Ethical acceptability

1 = Major ethical problems

2 = Minor ethical problems

3 = No ethical problems

N.B. The above rating should be based on the existing data and not on mere assumptions.

Exercises

1. In a certain district (population, 150,000), sanitary conditions are very poor (only 5% of households have latrines) and diseases connected with poor sanitation, such as, gastroenteritis and worms are very common. The Ministry of Health has initiated a sanitation project that aims at increasing the number of households with latrines by 20% each year. The project provides materials and the population should provide labour. Two years later, less than half of the target has been reached.

   State the discrepancy, research question and the possible answers. Is this problem situation researchable?

2. Go to the nearby health institution and identify three health problems. Discuss about these health problems and rate them based on the selection criteria.

   When rating these problems based on the criteria, use the rating scale indicated at the bottom of the table (you can also refer to the "Scales for rating research topics" presented in section 2.4b). You can do the exercise in small groups.

   Which topic do you select for research? Defend your first choice in a plenary session.

Rating Sheet

Criteria for selecting a research topic	Proposed topic		
	Health problem **I**	Health problem **II**	Health problem **III**
Relevance			
Avoidance of duplication			
Feasibility			
Political acceptability			
Applicability			
Urgency of data needed			
Ethical acceptability			
Total			

Rating scale: 1 = low, 2 = medium, 3 = high

CHAPTER THREE

ANALYSIS AND STATEMENT OF THE PROBLEM

3.1 Learning objectives

After completing this chapter, the student should be able to:

1. Describe the advantages of a systematic analysis of a problem
2. Describe the importance of a clear statement of a problem
3. Enumerate the points that should be included in the statement of a problem

3.2 Introduction

Was the problem adequately analysed to include all possible contributory factors from different sectors? Was it clearly stated? These questions should be clearly answered before trying to develop the research proposal. The sections that follow are devoted to giving the principles useful for the analysis and statement of the given problem.

3.3 Analyzing the problem

A systematic analysis of the problem, completed jointly by the researchers, health workers, managers, and community representatives is a very crucial step in designing the research because it:

- Enables those concerned to ***bring together*** their knowledge of the problem,
- ***Clarifies*** the ***problem*** and the possible ***factors*** that may be contributing to it,
- ***Facilitates decisions*** concerning the focus and scope of the research.

3.4 Formulating the problem statement

After identifying, selecting and analyzing the problem, the next major section in a research proposal is “statement of the problem”

a) Why is it important to state and define the problem well?

Because a clear statement of the problem:

- Is the ***foundation*** for the further development of the research proposal (research objectives, methodology, work plan, etc);
- Makes it easier to find information and reports of similar studies from which your own study design can benefit;
- Enables the researcher to systematically point out why the proposed research on the problem should be undertaken and what you hope to achieve with the study results.

b) Points that need to be considered for justifying the selected research problem

A health problem selected to be studied has to be justified in terms of its:

- Being a current and existing problem which needs solution
- Being a widely spread problem affecting a target population
- Effects on the health service programmes
- Being a problem which concerns the planners, policy makers and the communities at large.

c) Information included in the statement of a problem

- A **brief description** of socioeconomic and cultural characteristics and an overview of health status.
- A more detailed description of the nature of the problem
  - basic description of the research problem
  - the discrepancy between what is and what should be
  - its size, distribution, and severity (who is affected, where, since when, etc.)

- An analysis of the major factors that may influence the problem and a convincing argument that available knowledge is insufficient to answer a certain question and to update the previous knowledge.

- A brief description of any solutions that have been tried in the past, how well they have worked, and why further research is needed.

- A description of the type of information expected to result from the project and how this information will be used to help solve the problem

- If necessary, a short list of definitions of crucial concepts used in the statement of the problem.

A list of abbreviations may be annexed to the proposal, but each abbreviation also has to be written out in full when introduced in the text the first time.

3.5 Exercises

1. Why do we need to analyze the research problem?
2. What are the points required to justify the selected research problem?
3. What information should be included in the statement of a problem?

CHAPTER FOUR

LITERATURE REVIEW

4.1 Learning objectives

After completing this chapter, the student should be able to:

1. Describe the reasons for reviewing available literature and other information during the preparation of a research proposal.
2. Describe the resources that are available for carrying out such a review.
3. Record (organize) information obtained from literature on an index card.

4.2 Introduction

At the outset of his/her study the investigator should be acquainted with the relevant literature. It is of **minimal use** to wait until a report is written.

4.3 Use of literature review

- It prevents you from duplicating work that has been done before.
- It increases your knowledge on the problem you want to study and this may assist you in refining your "statement of the problem".
- It gives you confidence why your particular research project is needed.
- To be familiar with different research methods

4.4 Sources of information

- Card catalogues of books in libraries
- Organizations (institutions)
- Published information (books, journals, etc.)
- Unpublished documents (studies in related fields, reports, etc.)
- Computer based literature searches such as Medline
- Opinions, beliefs of key persons

Some examples of resources where information could be obtained are:

- Clinic and hospital based data from routine activity statistics
- Local surveys, annual reports
- Scientific conferences
- Statistics issued at region and district levels
- Articles from national and international journals (e.g., The Ethiopian Journal of Health Development, The Ethiopian Medical Journal, The East African Medical journal, The Lancet, etc.)
- Internet
- Documentation, reports, and raw data from the Ministry of Health, Central Statistical Offices, Nongovernmental organizations, etc.

References that are identified:

- Should first be skimmed or read
- Then summaries of the important information in each of the references may be recorded on separate index cards
- These should then be classified so that the information can easily be retrieved

4.5 Organization of information on index cards

The index cards should contain:

- Key words
- A summary of the contents of books or articles which is relevant to one's own study
- A brief analysis of the content, with comments such as:
  - how information from that particular study could be used in one's own study
- Information obtained from key persons could also be summarized on the index card

After collecting the required information on index cards, the investigator should decide in which order he/she wants to discuss previous research findings:

- from global to local
- from broader to focused

- from past to current

In conclusion, while reviewing a literature, all what is known about the study topic should be summarized with the relevant references. This review should answer

- ***How much is known?***
- ***What is not known?***
- ***What should be done based on what is lacking?***

Overall, the literature review should be adequate, relevant and critical. In addition to this, appropriate referencing procedures should always be followed in research proposals as well as in research reports. While reviewing a literature give emphasis to both positive and negative findings and avoid any distortion of information to suit your own study objectives.

Finally, after an exhaustive literature review, summarize the findings and write a coherent discussion by indicating the research gap which supports the undertaking of your study.

4.6 Exercises

1. Why is literature review important when preparing a proposal?
2. The presentation of research results or scientific publications from other writers without quoting the author is not appropriate. Does this statement sound true? Justify your answer.
3. Mention some of the sources of information in your area and describe how such information could be summarized on index cards.

CHAPTER FIVE

OBJECTIVES

5.1 Learning objectives

After completing this chapter, the student should be able to:

1. Describe the need for the development of research objectives
2. Differentiate between general and specific objectives
3. Formulate specific objectives and hypotheses

5.2 Introduction

Having decided what to study, and knowing why s/he wants to study it, the investigator can now formulate his study objectives. Objectives should be closely related to the statement of the problem. For example, if the problem identified is low utilization of health stations in a rural district, the general objective of the study could be to assess the reasons for this low utilization. If we break down this general objective into smaller and logically connected parts, then we get specific objectives.

5.3 Definitions

General objectives: aim of the study in general terms

Example: In a study on missed opportunities for EPI in Addis Ababa the general objective was: ***"to assess missed opportunities for EPI in Addis Ababa".***

Specific objectives: measurable statements on the specific questions to be answered. Unlike the general objectives, the specific objectives are more specific and are related to the research ***problem situation.*** They indicate the ***variable to be examined and measured.***

Example: In the study of missed opportunity for EPI in Addis Ababa the specific objectives could be:

- To find out the magnitude of missed opportunities for children who attend OPD, MCH, CDD, etc. in Addis Ababa,

- To examine the reasons for children not being immunized while attending the OPD, MCH, CDD, etc. services.

5.4 Formulation of the research objectives

The formulation of objectives will help us to:

- Focus the study (narrowing it down to essentials)
- Avoid collection of data that are not strictly necessary for understanding and solving the identified problem
- Organize the study in clearly defined parts

The explicit formulation of study objectives is an essential step in the planning of a study. It is said that “a question well-stated is a question half-answered”, but a question that is poorly stated or unstated is unlikely to be answered at all.

How should we state our objectives?

We have to make sure that our objectives:

- Cover the different aspects of the problem and its contributing factors in a **coherent way** and in a **logical sequence**
- Are **clearly expressed** in measurable terms
- Are **realistic** considering local conditions
- Meet the purpose of the study
- Use **action verbs** that are specific enough to be measured

Examples of action verbs are:

- to determine
- to compare
- to verify

- to calculate
- to describe
- to find out
- to establish

Avoid the use of vague non-action verbs such as;

- to appreciate
- to understand
- to study
- to believe

Research objectives can be stated as:

- **Questions** - the objectives of this study are to answer the following questions ….
- **Positive sentence** - the objectives of this study are to find out, to establish, to determine, …
- **Hypothesis** - the objective of this study is to verify the following hypothesis (examples are given below)

Based on the type of the study problem, it might be possible to develop explanations for the problem that can be tested. If so, we can formulate hypotheses in addition to the other study objectives.

A hypothesis is a prediction of a relationship between one or more variables and the problem under study. That is, It specifies the relationship among variables. These variables are to be statistically tested at a later stage. In order to measure the relationship among variables to be studied the dependent and independent variables need to be identified. A few examples are given below:

1. The health of children living in rural villagization projects is better than those living in traditional rural communities.
2. To examine whether there is any significant difference between district “A” and district “B” with respect to their malaria prevalence rates

3. An increase in the frequency of face washing is followed by a reduction in trachoma prevalence

One of the most important problems usually observed among students is the tendency of stating too many study objectives which are not appropriately addressed (or sometimes will be forgotten) in the sections that follow. It should be noted that it is on the bases of these specific objectives that the methods, results and discussion sections will be presented. For example, sample size calculations for each stated objective and identifying (selecting) the most appropriate sample size that will answer the required research questions is not covered in the development of most research proposals. This is also true during the write up of the completed research work. It is not uncommon to come across a situation in which some of the specific objectives are not addressed in the results section at all. It is therefore advisable to limit the number of specific objectives. In most practical situations, the number of specific objectives should not exceed three.

5.5 Exercises

1. Define general objectives, specific objectives and hypotheses by giving your own examples.
2. The objectives of a study should be written after the statement of the research problem and before the methods section. Does this statement sound true? Justify your answer.
3. List the characteristics of research objectives.
4. Comment on the statement: “A question well-stated is a question half-answered”.
5. Mention some of the problems that may arise as a result of having too many objectives.

CHAPTER SIX
RESEARCH METHODS

6.1 Learning objectives

After completing this chapter, the student should be able to:

1. Identify the pertinent questions to consider when developing the methodology of a research proposal
2. Describe and understand the various components of the methods section in a research proposal
3. Explain the cyclical nature of the different steps in designing the methodology.

6.2 Introduction

In the previous chapters we have dealt with the identification, selection, analysis and statement of the problem. The importance of literature review and formulation of study objectives were also emphasized. Now we must decide exactly how we are going to achieve our stated objectives. That is, what new data do we need to shed light on the problem we have selected and how we are going to collect and process these data. The major issues that constitute the "methods section" of a research proposal will be dealt in the sections that follow.

6.3 Types of study designs

A study design is the process that guides researchers on how to collect, analyze and interpret observations. It is a **l*ogical model*** that guides the investigator in the various stages of the research.

Several classifications of study types are possible, depending on what research strategies are used.

1. **Non-intervention (Observational) studies** in which the researcher just observes and analyses researchable objects or situations but does not intervene; and

2. **Intervention studies** in which the researcher manipulates objects or situations and measures the outcome of his manipulations (e.g., by implementing intensive health education and measuring the improvement in immunisation rates.)

Study designs could be exploratory, descriptive or analytical

1. Exploratory studies

An **exploratory study** is a small-scale study of relatively short duration, which is carried out when little is known about a situation or a problem. It may include description as well as comparison.

For example:

> A national AIDS Control Programme wishes to establish counseling services for HIV positive and AIDS patients, but lacks information on specific needs patients have for support. To explore these needs, a number of in-depth interviews are held with various categories of patients (males, females, married and single) and with some counselors working on a programme that is already under way.

When doing exploratory studies we *describe* the needs of various categories of patients and the possibilities for action. We may want to go further and try to explain the differences we observe (e.g., in the needs of male and female AIDS patients) or to identify causes of problems. Then we will need to *compare* groups.

> **If the problem and its contributing factors are not well defined it is always advisable to do an exploratory study before embarking on a large-scale descriptive or comparative study.**

2. Descriptive studies:

Descriptive studies may be defined as studies that describe the patterns of disease occurrence and other health-related conditions by ***person place*** and ***time.***

Personal variables include: basic demographic factors, such as age, sex marital status or occupation, as well as the consumption of various types of food or medication use.

Characteristics of place refer to the **geographic distribution of disease,** including variation among countries or within countries, such as between urban and rural areas.

With regard to time, descriptive studies may examine **seasonal patterns in disease onset, etc.**

Uses of descriptive studies

- They can be done fairly quickly and easily.
- Allow planners and administrators to allocate resources
- Provide the first important clues about possible determinants of a disease (useful for the formulation of hypotheses)

Types of descriptive studies

a) Case reports and case series

Case report: a careful, detailed report by one or more clinicians of the ***profile of a single patient.***

The individual case report can be expanded to a ***case series,*** which describes characteristics of a number of patients with a given disease.

Uses

- Important link between clinical medicine and epidemiology
- One of the first steps in outbreak investigation
- Often useful for hypothesis generating and examining new diseases, but conclusions about etiology cannot be made.

b) Ecological studies: data from entire populations are used to compare disease frequencies between different groups during the same period of time or in the same population at different points in time.

Example: Countries with low cigarette consumption have lower lung cancer rates than those countries with high cigarette consumption.

- Ecological studies are usually quick and easy to do and can be done with already available information.

- Since ecological studies refer to whole populations rather than to individuals, it is not possible to link an exposure to occurrence of disease in the same person.

c) Cross-sectional studies

A cross-sectional (prevalence) study provides information concerning the situation at a given time. In this type of study, the status of an individual with respect to the presence or absence of both exposure and disease is assessed at the same point in time.

- Usually involve collection of new data.
- In general, measure prevalence rather than incidence
- Not good for studying rare diseases or diseases with short duration; also not ideal for studying rare exposures.

For factors that remain unaltered over time, such as sex, blood group, etc., the cross-sectional survey can provide evidence of a valid statistical association.

As can be noted from the above explanation, a cross-sectional study can be either analytical or descriptive, according to its purpose. If data are collected both on exposures and outcomes of interest, and if the data are analysed so as to demonstrate differences either between exposed and non-exposed groups, with respect to the outcome, or between those with the outcome and those without the outcome, with respect to the exposure, then this is an analytical cross-sectional study. If the information collected is purely of a descriptive nature, not involving the **comparison** of groups formed on the basis of exposure or outcome status, then this is a descriptive cross-sectional study. Often a cross-sectional study may have both descriptive and analytical components.

Nowadays, there is an increasing emphasis on the value of longitudinal studies in which observations are repeated in the same community over a prolonged period (i.e., longitudinal

studies provide the required data at **more than one point** in time unlike cross- sectional surveys)**.**

II. Analytic studies

Analytic studies may be defined as studies used ***to test hypotheses*** concerning the relationship between a suspected risk factor and an ***outcome*** and to **measure the magnitude of the association** and its ***statistical significance.***

Analytic study designs can be divided into two broad design strategies: Observational and intervention.

Observational studies

- No human intervention involved in assigning study groups; simply observe the relationship between exposure and disease.
- Subject to many potential biases, but by careful design and analysis, many of these biases can be minimized.
- Examples of observational studies: comparative cross-sectional, cohort and case-control studies.

a) Comparative cross-sectional studies: Depending on the purpose of a given study, a cross-sectional survey could have an analytical component (see section 6.3, 2c, above).

b) Cohort studies: Study groups identified by exposure status prior to ascertainment of their disease status and both exposed and unexposed groups followed in identical manner until they develop the disease under study, they die, the study ends, or they are lost to follow-up.

Strengths and limitations of the cohort study design

Strengths:

- Is of particular value when the **exposure is rare**
- Can examine **multiple effects** of a single exposure
- Allows direct measurement of **incidence** of disease in the exposed and non-exposed groups.

Limitations:

- Is inefficient for the evaluation of rare diseases
- Expensive and time consuming
- Validity of the results can be seriously affected by losses to follow-up.

c) Case-control studies: Group of subjects with the disease (cases) and group of subjects without the disease (controls) are identified. Information, about previous exposures are obtained for cases and controls, and frequency of exposure compared for the two groups.

Strengths and limitations of the case-control study design

Strengths:

- Is relatively quick and inexpensive
- Is optimal for the evaluation of rare diseases.
- Can examine multiple etiologic factors for a single disease.

Limitations:

- Is inefficient for the evaluation of rare exposures
- Cannot directly compute incidence rates of disease in exposed and non- exposed individuals.
- Is particularly prone to bias compared with other analytic designs, in particular, selection and recall bias.

Intervention studies

In intervention studies, the researcher manipulates a situation and measures the effects of this manipulation. Usually (but not always) two groups are compared, one group in which the intervention takes place (e.g. treatment with a certain drug) and another group that remains 'untouched' (e.g. treatment with a placebo).

The two categories of intervention studies are:

- experimental studies and
- quasi-experimental studies

1. Experimental studies

An experimental design is a study design that gives the most reliable **proof for causation**. In an **experimental study,** individuals are randomly allocated to at least two groups. One group is subject to an intervention, or experiment, while the other group(s) is not. The outcome of the intervention (effect of the intervention on the dependent variable/problem) is obtained by comparing the two groups. A number of experimental study designs have been developed. These are widely used in laboratory settings and in clinical settings. For ethical reasons, the opportunities for experiments involving human subjects are restricted. However, randomised control trials of new drugs are common.

At community level, where health research is frequently undertaken, we experience not only ethical but also practical problems in carrying out experimental studies. In real life settings, it is often impossible to assign persons at random to two groups, or to maintain a control group. Therefore, experimental research designs may have to be replaced by quasi-experimental designs.

2. Quasi-experimental studies

In a **quasi-experimental study**, one characteristic of a true experiment is missing, either randomisation or the use of a separate control group. A quasi-experimental study, however, always includes the manipulation of an independent variable which is the intervention.

One of the most common quasi-experimental designs uses two (or more) groups, one of which serves as a control group in which no intervention takes place. Both groups are observed before as well as after the intervention, to test if the intervention has made any difference. (This quasi-experimental design is called the 'non-equivalent control group design' because the subjects in the two groups (study and control groups) have not been randomly assigned.)

Another type of design that is often chosen because it is quite easy to set up uses only **one group** in which an intervention is carried out. The situation is analysed before and after the intervention to test if there is any difference in the observed problem. This is called a 'BEFORE-AFTER' study. This design is considered a 'pre-experimental' design rather than a 'quasi-experimental' design because it involves neither randomisation nor the use of a control group.

Intervention (experimental) studies can also be considered either therapeutic or preventive.

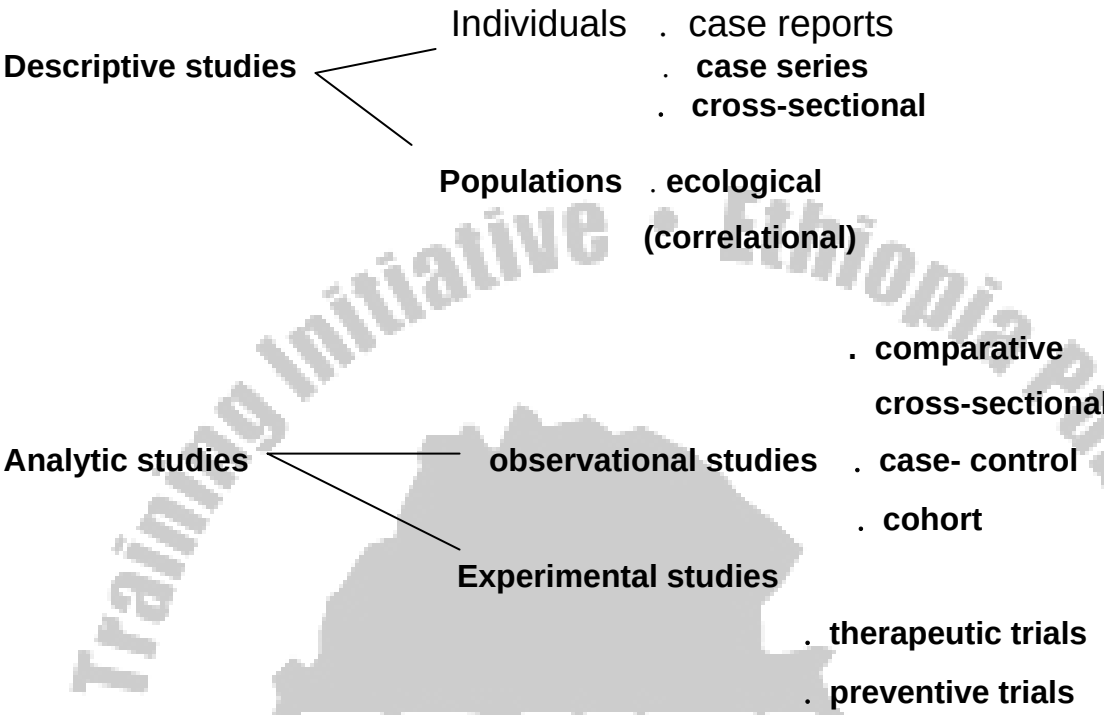
Therapeutic trials are conducted **among patients with a particular disease** to determine the ability of an agent or procedure to diminish symptoms, prevent recurrence, or decrease risk of death from that disease.

A preventive trial (community trial) involves the evaluation of whether an agent or ***procedure reduces the risk of developing disease*** among those ***free from that*** condition at enrolment. Thus, preventive trials can be conducted among individuals at usual risk (e.g. vaccine trials)

A particular research question may be addressed using different approaches. The choice of study design for investigation is influenced by:

- Particular features of the exposure and disease.
- Logistic considerations of available resources.
- Results from previous studies and gaps in knowledge that remain to be filled.
- Ingenuity and creativity of the researcher

Summary



6.4 Study population

At an early stage in the planning of any investigation decisions must be made concerning the study population. That is, concerning the population of individual units (whether they are persons, households, etc.) to be investigated. The population under consideration should be clearly and explicitly defined in terms of place, time, and other relevant criteria. If the study population comprises cases of a disease the procedures to be used for case identification should be stated. If controls are to be chosen their method of selection should be stated.

Often the investigator will have implicitly chosen his study population when he defined the topic of his investigation, by reason of his interest in a specific community or a specific health program.

In other instances, particularly when an analytic survey or an experiment is being planned, the investigator may require purposively to select a study population. In so doing he must consider questions of appropriateness and practicability.

The appropriateness of the study population refers to its suitability for the attainment of the objectives of the study.

The selection of study population on the basis of suitability usually affects the validity of subsequent generalizations from the findings. This situation requires a close attention at the early stage of the given study. Two examples are given below.

a) **Volunteer populations**: Persons who volunteer to enter a study may differ in many respects from those who do not so volunteer, and therefore the findings in a volunteer population do not necessarily apply to the population at large.

b) **Hospital or clinic populations**: Persons receiving medical care are obviously not representative of the general population from which they have come from. That is, persons treated in hospital for a certain disease may differ from those patients with the same disease but not receiving care for it.

Practical questions such as the following could also arise.

- Is the proposed population the one that would give the required information?
- Will the population cooperate to participate in the study, or will it be a 'resistant' one?
- If it is proposed to study patients with a specific disease, will it be possible to identify enough cases to yield useful conclusions?
- If a long term 'follow up' study is planned, is the population so mobile that it may be difficult to maintain contact with the subjects?

A preliminary exploratory study may sometimes be required in order to answer such questions.

6.5 Operational Definitions of Variables

Before we directly go to the operational definition of variables it would be important to discuss about the nature of variables first.

Definition: A variable is a characteristic of a person, object, or phenomenon that can take on different values.

A simple example of a variable is a person's age. The variable can take on different values, such as, 20 years old, 30 years old, and so on. Other examples of variables are:

a) weight in kilograms
b) height in centimeters
c) monthly income in Birr
d) marital status (single, married, divorced and widowed)
e) job satisfaction index (1 to 5)
f) occupation (civil servant, farmer, student, et.)
g) disease condition (presence or absence of a disease)

The first three variables (a to c) are **numerical** variables because they are expressed in numbers (metric data). Since the values of the remaining three variables (d to g) are expressed in categories, we call them **categorical** variables.

Because in the health research we often look for associations, it is important to make a distinction between **dependent and independent** variables. Both the dependent and independent variables together with their operational definitions (when necessary) should be stated.

Definitions:
The variable that is used to describe or measure the problem under study is called the dependent variable. The variables that are used to describe or measure the factors that are assumed to influence (or cause) the problem are called independent variables.

For example, in a study of relationship between smoking and lung cancer, "suffering from lung cancer" (with the values yes, no) would be the **dependent** variable and "smoking" (with the values no, less than a packet/day, 1 to 2 packets/day, more than 2 packets/day) would be the **independent** variable.

Background variables - In almost every study involving human subjects, background variables, such as, age, sex, educational status, monthly family income, marital status and

religion will be included. These background variables are often related to a number of independent variables, so that they influence the problem indirectly. Hence they are called background variables or background characteristics.

Confounding variable - A variable that is associated with the problem and with a possible cause of the problem is a potential confounding variable. This type of variable may either strengthen or weaken the apparent relationship between the problem and a possible cause.

Composite variable - A variable based on two or more other variables may be termed a composite variable. Incidence and prevalence rates, sex ratios, and other rates and ratios are composite variables, since they are based on separate numerator and denominator information.

I. Operationalising variables by choosing appropriate indicators

Note that the different values of many of the variables presented above can easily be determined. However, for some variables it is sometimes not possible to find meaningful categories unless the variables are made operational with one or more precise **INDICATORS**. Operationalising variables means that you make them ‘measurable'.

For example:

1. In a study on VCT acceptance, you want to determine the **level of knowledge** concerning HIV in order to find out to what extent the factor ‘poor knowledge’ influences willingness to be tested for HIV. The variable ‘level of knowledge’ cannot be measured as such. You would need to develop a series of questions to assess a person’s knowledge, for example on modes of transmission of HIV and its prevention methods. The answers to these questions form an **indicator** of someone’s knowledge on this issue, which can then be categorised. If 10 questions were asked, you might decide that the knowledge of those with:

- 0 to 3 correct answers is poor,
- 4 to 6 correct answers is reasonable, and
- 7 to 10 correct answers is good.

When defining variables on the basis of the problem analysis diagram, it is important to realise which variables are measurable as such and which ones need indicators. Once appropriate indicators have been identified we know exactly what information we are looking for. This makes the collection of data as well as the analysis more focused and efficient.

2. Nutritional status of under-5 year olds is another example of a variable that cannot be measured directly and for which you would need to choose appropriate indicators. Widely used indicators for nutritional status include weight for age, weight for height, height for age, and upper-arm circumference. For the classification of nutritional status, internationally accepted categories already exist, which are based on standard growth curves. For the indicator weight/age, for example, children are:

- Well nourished if they are above 80% of the standard
- Moderately malnourished if they are between 60% and 80%
- Severely malnourished if they are below 60%

II. Defining variables and indicators of variables

To ensure that everyone (the researcher, data collectors, and eventually the reader of the research report) understands exactly what has been **measured** and to ensure that there will be consistency in the **measurement,** it is necessary to clearly define the variables (and indicators of variables). For example, to define the indicator “waiting time” it is necessary to decide what will be considered the starting point of the “waiting period” e.g. Is it when the patient enters the front door, or when he has been registered and obtained his card?

For certain variables, it may not be possible to adequately define the variable or the indicator immediately because further information may be needed for this purpose. The researcher may need to review the literature to find out what definitions have been used by other researchers, so that he can standardize his definitions and thus be able later to easily compare his findings with those of the other studies. In some cases the opinions of “experts” or of community members of health care providers may be needed in order to define the variable or indicator.

The variables to be studied are selected on the basis of their relevance to the objectives of the investigation.

- **The initial list is usually too long**
- **It has to be pruned to facilitate the collection and processing of the data.**

Once the variables are selected, each of them should be clarified. There are two aspects to be considered.

1. Clear definition of variables in terms of **objectively measurable facts** (i.e., operational definition) - this was repeatedly mentioned (addressed) in the above examples
2. The **scale of measurement** to be used in data collection.

Unless the variables are clearly and explicitly defined, there can be no assurance that, if the study is performed by a different investigator, or repeated by the same investigator, similar findings would be obtained.

The following example shows the different definitions (two different definitions) given to "obesity".

The two kinds of definitions are: conceptual and operational. The conceptual definition is often akin to a dictionary definition.

e.g. " Obesity" may be defined as:

" excessive fatness", "overweight", etc.

In contrast, the operational definition is heavily influenced by considerations of practicability. " Obesity", for example, might be operationally defined as: " a weight, based on weighing in underclothes and without shoes, which exceeds, by 10% or more, the mean weight of persons of the subject's sex, age and height (in a specified population at a specified time)".

In general, operational definitions of variables are used in order to:

- **Avoid ambiguity**
- **Make the variables to be more measurable**

Scales of Measurement

As part of the process of clarifying each of the variables to be studied, its scale of measurement should be specified. There are four types of scales of measurement: **Nominal, Ordinal, Interval and Ratio.** They are listed in ascending order of power and preference.

1. **Nominal Scale:** This consists of two or more named categories (classes) which are qualitatively different from each other.

   E.g Sex: male (1); Female (2)

   Marital status: 1. Married 2. Single 3. Divorced 4. Widowed

2. **Ordinal scale:** This has the additional quality that the categories are ranked and have implied order. However, the intervals between classes are not necessarily equal.

   Example 1. Severity of a disease: Severe (grade III); moderate (grade II); mild (grade I); absent (grade 0).

   Example 2. Educational status: 0; 1-6; 7-8; 9 -12; more than 12.

3. **Interval scale:** This has the additional quality that the intervals between classes are equal.

   Example: Temperature (in Celsius)

   Equal differences between any pair of numbers in the scale indicate equal differences in the attribute being measured. The difference in temperature between 20… C and 25…C is the same as the difference between 30…C and 35…C. The ratio between numbers in the scale is not, however, necessarily the same as that between the amounts of the attribute. That is, a room at 30… C is not 'twice as hot' as one at 15…C. This is because *the zero on the scale does not indicate absence of the attribute.*

4. **Ratio scale:** This has the additional quality that zero indicates absence of the attribute. As a result, the ratio between numbers in the scale is the same as that between the amounts of the attribute being measured.

Example: Weight measured in kilograms, height in cms., etc.

6.6 Sampling

What is sampling?

Sampling involves the selection of a number of study units from a defined study population. The population is too large for us to consider collecting information from all its members. Instead we select a sample of individuals hoping that the sample is representative of the population.

When taking a sample, we will be confronted with the following questions:

a) What is the group of people from which we want to draw a sample?
b) How many people do we need in our sample?
c) How will these people be selected?

Definitions

Target population (reference population): Is that population about which an investigator wishes to draw a conclusion.

Study population (population sampled): Population from which the sample actually was drawn and about which a conclusion can be made. For Practical reasons the study population is often more limited than the target population. In some instances, the target population and the population sampled are identical.

Sampling unit: The unit of selection in the sampling process. For example, in a sample of districts, the sampling unit is a district; in a sample of persons, a person, etc.

Study unit: The unit on which the observations will be collected. For example, persons in a study of disease prevalence, or households, in a study of family size.

N.B. The sampling unit is not necessarily the same as the study unit.

Sample design: The scheme for selecting the sampling units from the study population.

Sampling frame: The list of units from which the sample is to be selected.

The existence of an adequate and up-to-date sampling frame often defines the study population.

Sampling methods

An important issue influencing the choice of the most appropriate sampling method is whether a sampling frame is available, that is, a listing of all the units that compose the study population.

a) **Non-probability sampling methods**

Examples:

1. Convenience sampling: is a method in which for convenience sake the study units that happen to be available at the time of data collection are selected.
2. Quota sampling: is a method that insures that a certain number of sample units from different categories with specific characteristics appear in the sample so that all these characteristics are represented. In this method the investigator interviews as many people in each category of study unit as he can find until he has filled his quota.
3. Purposeful sampling strategies for qualitative studies: Qualitative research methods are typically used when focusing on a limited number of informants, whom we select strategically so that their in-depth information will give optimal insight into an issue about which little is known. This is called purposeful sampling.

The above sampling methods do not claim to be representative of the entire population.

Random sampling strategies to collect quantitative data: If the aim of a study is to *measure* variables distributed in a population (e.g., diseases) or to *test hypotheses* about which factors are contributing significantly to a certain problem, we have to be sure that we can generalise the findings obtained from a sample to the total study population. Then, purposeful sampling methods are inadequate, and probability or random sampling methods have to be used.

b) Probability sampling methods: They involve random selection procedures to ensure that each unit of the sample is chosen on the basis of chance. All units of the study population should have an equal or at least a known chance of being included in the sample.

1. **Simple Random Sampling (SRS):** This is the most basic scheme of random sampling. To select a simple random sample you need to:

   - Make a numbered list of all the units in the population from which you want to draw a sample. Each unit on the list should be numbered in sequence from 1 to N (Where N is the Size of the population).
   - Decide on the size of the sample
   - Select the required number of sampling units, using a "lottery" method or a table of random numbers.

2. Systematic Sampling: Individuals are chosen at regular intervals (for example, every 5^{th}, 10^{th}, etc.) from the sampling frame. Ideally we randomly select a number to tell us where to start selecting individuals from the list. For example, a systematic sample is to be selected from 1000 students of a school. The sample size is decided to be 100. The sampling fraction is: 100/1000 = 1/10. The number of the first student to be included in the sample is chosen randomly by picking one out of the first ten pieces of paper, numbered 1 to 10. If number 5 is picked, every tenth student will be included in the sample, starting with student number 5, until 100 students are selected. Students with the following numbers will be included in the sample: 5,15, 25, 35,45, . . . , 985, 995.

- Systematic Sampling is usually less time consuming and easier to perform than SRS.
- It provides a good approximation to SRS.
- Should not be used if there is any sort of cyclic pattern in the ordering of the subjects on the list**.**
- Unlike SRS, systematic sampling can be conducted without a sampling frame (useful in some situations where a sampling frame is not readily available).

4. **Stratified sampling:** If it is important that the sample includes representative groups of study units with specific characteristics (for example, residents from urban and rural areas), then the sampling frame must be divided into groups, or strata, according to these characteristics. Random or systematic samples of a predetermined size will then have to be obtained from each group (stratum). This is called stratified sampling.

Some of the reasons for stratifying the population may be:

- Different sampling schemes may be used in different strata, e.g. Urban and rural
- Conditions may suggest that prevalence rates will vary between strata: the overall estimate for the whole population will be more precise if stratification is used.
- Administrative reasons may make it easier to carry out the survey through an organization with a regional structure.

5. **Cluster sampling:** When a list of groupings of study units is available (e.g. villages, etc.) or can be easily compiled, a number of these groupings can be randomly selected. The selection of **groups** of study units (clusters) instead of the selection of study units **individually** is called cluster sampling. Clusters are often geographic units (e.g. districts, villages) or organizational units (e.g. clinics).

6. **Multi-Stage Sampling:** This method is appropriate when the population is large and widely scattered. The number of **stages** of sampling is the number of times a sampling procedure is carried out.

    - The primary sampling unit **(PSU)** is the sampling unit (or unit of selection in the sampling procedure) in the **first sampling stage;**

- The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc.

e.g. After selection of a sample of clusters (e.g. household), further sampling of individuals may be carried out within each household selected. This constitutes two-stage sampling, with the PSU being households and the SSU being individuals.

Advantages: less costly, we only need to draw up a list of individuals in the clusters actually selected, and we can do that when we arrive there.

Disadvantage: less precise than SRS.

When we take a sample, our results will not exactly equal the correct results for the whole population. That is, our results will be subject to errors. This error has two components: sampling and non-sampling errors.

a) Sampling error (i.e., random error)

Random error, the opposite of reliability (i.e., Precision or repeatability), consists of random deviations from the true value, which can occur in **any** direction.

Sampling error (random error) can be minimized by increasing the size of the sample.

Reliability (or precision): This refers to the repeatability of a measure, i.e., the degree of closeness between repeated measurement of the same value. Reliability addresses the question, if the same thing is measured several times, how close are the measurements to each other?

The sources of variation resulting in poor reliability include:

a) Variation in the characteristic of the subject being measured. Example: blood pressure
b) The measuring instruments, e.g. questionnaires
c) The persons collecting the information (observer variation)

Inter-observer variation: differences between observers in measuring the same observation

Intra-observer variation: differences in measuring the same observation by the same observer on different occasions.

b) Non Sampling error (i.e., bias)

Bias, the opposite of validity, consists of systematic deviations from the true value, **always in the same direction.**

It is possible to eliminate or reduce the non-sampling error (bias) by **careful design** of the sampling procedure.

Validity: This refers to the degree of closeness between a measurement and the true value of what is being measured. Validity addresses the question, how close is the **measured value** to the **true value**?

To be accurate, a measuring device must be both valid and reliable. However, if one cannot have both, validity is more important in situations when we are interested in the absolute value of what is being measured. Reliability on the other hand is more important when it is not essential to know the absolute value, but rather we are interested in finding out if there is a trend, or to rank values.

Examples of types of bias in sampling include:

Bias resulting from incompleteness of the sampling frame: accessibility bias, seasonability bias, self-reporting bias, volunteer bias, non-response bias etc.

Non-response bias refers to failure to obtain information on some of the subjects included in the sample to be studied. It results in significant bias when the following two situations are both fulfilled.

1. When non-respondents constitute a significant proportion of the sample.
2. When non-respondents differ significantly from respondents.

The issue of non-response should be considered during the planning stage of the study:

a) Non-response should be kept to a minimum. E.g. below 15%

Methods that may help in maintaining non-response at a low level could be:

- Training data collectors to initiate contact with study subjects in a respectful way and convince them about the importance of the given study (this minimizes the refusal type of non-response)
- Offering incentives to encourage participation (this should be done by taking account of the potential problems that may arise in conducting future research)
- By making repeated attempts (at least 3 times) to contact study subjects who were absent at the time of the initial visit.

b) The number of non-responses should be documented according to type, so as to facilitate an assessment of the extent of bias introduced by non-response.

c) As much information as possible should be collected on non-respondents, so as to see in what ways they may differ from respondents.

- Selection bias cannot be corrected by increasing the size of the sample, **why?** How do you remove this type of bias?

6.7 Sample size determination

- In planning any investigation we must decide how many people need to be studied in order to answer the study objectives. If the study is too small we may fail to detect important effects, or may estimate effects too imprecisely. If the study is too large then we will waste resources.

- In general, it is much better to increase the accuracy of data collection (by improving the training of data collectors and data collection tools) than to increase the sample size **after a certain point.**

- The eventual sample size is usually a compromise between what is desirable and what is feasible.

- The feasible sample size is determined by the availability of resources. It is also important to remember that resources are not only needed to collect the information, but also to analyze it.

In order to calculate the required sample size, you need to know the following facts:

a) The reasonable estimate of the key proportion to be studied. If you cannot guess the proportion, take it as 50%.

b) The degree of accuracy required. That is, the allowed deviation from the true proportion in the population as a whole. It can be within 1% or 5%, etc.

c) The confidence level required, usually specified as 95%.

d) The size of the population that the sample is to represent. If it is more than 10,000 the precise magnitude is not likely to be very

important; but if the population is less than 10,000 then a smaller sample size may be required.

e) The difference between the two sub-groups and the value of the likelihood or the power that helps in finding a statistically significant difference.

- Note that 'e' is required when there are two population groups and the interest is to compare between two means or proportions.

Estimating a proportion

- Estimate how big the proportion might be (P)
- Choose the margin of error you will allow in the estimate of the proportion (say $\pm$ w)
- Choose the level of confidence that the proportion in the whole population is indeed between (p-w) and (p+w). We can never be 100% sure. Do you want to be 95% sure?
- The minimum sample size required, for a very large population (N>10,000) is:

$$n = Z^2 \, p(1-p) / w^2$$

Example 1 (Prevalence of diarrhoea)

a) p = 0.26 , w = 0.03 , Z = 1.96 (i.e., for a 95% C.I.)

$$n = (1.96)^2 \, (.26 \times .74) / (.03)^2 = 821.25 \approx 822$$

Thus, the study should include at least 822 subjects.

b) If the above sample is to be taken from a relatively small population (say N = 3000), the required minimum sample will be obtained from the above estimate by making some adjustment.

821.25 / (1+ (821.25/3000)) = 644.7 ≈ **645 subjects**

Example 2

A hospital administrator wishes to know what proportion of discharged patients are unhappy with the care received during hospitalization. If 95% Confidence interval is desired to estimate the proportion within 5%, how large a sample should be drawn?

$n = Z^2 p(1-p)/w^2 = (1.96)^2(.5\times.5)/(.05)^2 = 384.2 \approx 385$ patients

N.B. If you don't have any information about P, take it as 50% and get the maximum value of PQ which is 1/4 (i.e., 25%).

Estimating a mean

The same approach is used but with SE = $\sigma / \sqrt{n}$

The required (minimum) sample size for a very large population is given by :

$$\mathbf{n = Z^2 \sigma^2 / w^2}$$

Example: A health officer wishes to estimate the mean serum cholesterol in a population of men. From previous similar studies a standard deviation of 40 mg/100ml was reported. If he is willing to tolerate a marginal error of up to 5 mg/100ml in his estimate, how many subjects should be included in his study?

(α =5%, two sided)

a) If the population size is assumed to be very large, the required sample size would be:

$n = (1.96)^2 (40)^2 / (5)^2 = 245.86 \approx 246$ persons

- If the population size is, say, 2000, the required sample size would be 220 persons.

b) If the investigator anticipates that 15% of the subjects will fail to comply with the intended study, the sample size required will be:

n = (1/(1-0.15)) x 246 = 290 men

NB: σ^2 can be estimated from previous similar studies or could be obtained by conducting a small pilot study.

Comparison of two proportions

n (in each region) = $(p_1q_1 + p_2q_2)\ (f(\alpha,\beta))\ /\ ((p_1 - p_2)^2$

α = type I error (level of significance)
β = type II error (1-β = power of the study)

power = the probability of getting a significant result

f (α,β) =10.5, when the power = 90% and the level of significance = 5%

Eg. The proportion of nurses leaving the health service is compared between two regions. In one region 30% of nurses is estimated to leave the service within 3 years of graduation. In other region it is probably 15%.

Solution

The required sample to show, with a 90% likelihood (power), that the percentage of nurses is different in these two regions would be:
(assume a confidence level of 95%)

$n = (1.28+1.96)^2\ ((.3\times.7) +(.15\times.85))\ /\ (.30 - .15)^2 = 158$

158 nurses are required in each region

Comparison of two means (sample size in each group)

$n = (s_1^2 + s_2^2)\ f(\alpha,\beta)\ /\ (m_1 - m_2)^2$

m_1 and s_1^2 are mean and variance of group 1 respectively.
m_2 and s_2^2 are mean and variance of group 2 respectively.

N.B. Sample size calculation using the STATCALC calculator of the Epi Info program is given in chapter 8.

6.8 Plan for data collection

Why should you develop a plan for data collection?

A plan for data collection should be developed so that:

- you will have a clear overview of what tasks have to be carried out, who should perform them, and the duration of these tasks;
- you can organise both human and material resources for data collection in the most efficient way; and
- you can minimise errors and delays which may result from lack of planning (for example, the population not being available or data forms being misplaced).

It is likely that while developing a plan for data collection you will identify problems (such as limited manpower), which will require modification of the proposal. Such modifications might include adjustment of the sample size or extension of the period for data collection.

Stages in the Data Collection Process

Three main stages can be distinguished:
Stage 1: Permission to proceed
Stage 2: Data collection
Stage 3: Data handling

Stage 1: permission to proceed

Consent must be obtained from the relevant authorities, individuals and the community in which the project is to be carried out. This may involve organizing meetings at national or provincial level, at district and at village level. For clinical studies this may also involve obtaining written informed consent.

Stage 2: Data collection

When collecting our data, we have to consider:

- Logistics: who will collect what, when and with what resources
- Quality control

I. Logistics of data collection

WHO will collect WHAT data?

When allocating tasks for data collection, it is recommended that you first list them. Then you may identify who could best implement each of the tasks. If it is clear beforehand that your research team will not be able to carry out the entire study by itself, you might plan to look for research assistants to assist in relatively simple but time-consuming tasks.

HOW LONG will it take to collect the data for each component of the study?

Step 1: Consider:

- The time required to reach the study area(s);
- The time required to locate the study units (persons, groups, records); If you have to search for specific informants (e.g., users or defaulters of a specific service) it might take more time to locate informants than to interview them.
- The number of visits required per study unit. For some studies it may be necessary to visit informants a number of times, for example if the information needed is sensitive and can only be collected after informants are comfortable with the investigator or if observations have to be made more than once (for example, follow-up of pregnant mothers or malnourished children). Time needed for follow-up of non-respondents should also be considered.

Step 2: Calculate the number of interviews that can be carried out per person per day

Step 3: Calculate the number of days needed to carry out the interviews. For example:

- you need to do 200 interviews,
- your research team of 5 people can do 5 x 4 = 20 interviews per day,
- you will need 200:20 = 10 days for the interviews.

Step 4: Calculate the time needed for the other parts of the study, (for example, 10 days)

Step 5: Determine how much time you can devote to the study.

If the team has fewer days for fieldwork than the required, they would need additional research assistants to help complete this part of the study.

Note:

It is always advisable to slightly overestimate the period needed for data collection to allow for unforeseen delays.

WHEN should the data be collected?

The type of data to be collected and the demands of the project will determine the actual time needed for the data to be collected. Consideration should be given to:

- availability of research team members and research assistants,
- the appropriate season(s) to conduct the field work (if the problem is season-related or if data collection would be difficult during certain periods),
- accessibility and availability of the sampled population, and
- public holidays and vacation periods.

II. Ensuring quality

It is extremely important that the data we collect are of good quality, that is, reliable and valid. Otherwise we will come up with false or misleading conclusions.

Measures to help ensure good quality of data:

- **Prepare a field work manual for the research team as a whole,** including:

  — Guidelines on **sampling procedures** and what to do if respondents are not available or refuse to co-operate,

  — A clear **explanation** of the purpose and procedures of the study which should be used to introduce each interview, and

  — **Instruction sheets** on how to ask certain questions and how to record the answers.

- **Select your research assistants, if required, with care.** Choose assistants that are:

  — from the same educational level;

— knowledgeable concerning the topic and local conditions;

— not the object of study themselves; and

— not biased concerning the topic (for example, health staff are usually not the best possible interviewers for a study on alternative health practices).

- **Train research assistants carefully in all topics covered in the field work manual as well as in interview techniques** and make sure that all members of the research team master interview techniques such as:

  — asking questions in a neutral manner;

  — not showing by words or expression what answers one expects;

  — not showing agreement, disagreement or surprise; and

  — recording the answers precisely as they are provided, without sifting or interpreting them.

- **Pre-test research instruments and research procedures** with the whole research team, including research assistants.
- **Take care that research assistants are not placed under too much stress** (requiring too many interviews a day; paying per interview instead of per day).
- **Arrange for on-going supervision** of research assistants. If, in case of a larger survey, special supervisors have to be appointed, guidelines should be developed for supervisory tasks.
- **Devise methods to assure the quality** of data collected by all members of the research team. For example, quality can be assured by:

  — requiring interviewers to check whether the questionnaire is filled in completely before finishing each interview;

— asking the supervisor to check at the end of each day during the data collection period whether the questionnaires are filled in completely and whether the recorded information makes sense; and

— having the researchers review the data during the data analysis stage to check whether data are complete and consistent.

Stage 3: DATA HANDLING

Once the data have been collected and checked for completeness and accuracy, a clear procedure should be developed for handling and storing them. Decide if the questionnaires are to be numbered; identify the person who will be responsible for storing the data; and how they are going to be stored.

6.9 Methods of data collection

Having decided on how to design the research study, the next methodological design is how to collect information. The most commonly used methods of collecting information (quantitative data) are the use of documentary sources, interviews and self-administered questionnaires.

The choice of methods of data collection is based on:

- The accuracy of information they will yield
- Practical considerations, such as, the need for personnel, time, equipment and other facilities, in relation to what is available.

Accuracy and "practicability" are often inversely correlated. A method providing more satisfactory information will often be a more expensive or inconvenient one. Therefore, accuracy must be balanced against practical considerations (resources and other practical limitations)

The use of documentary sources

Clinical records and other personal records, death certificates, published mortality statistics, census publications, etc.

Advantages:

- Documents can provide ready made information relatively easily
- The best means of studying past events.

Disadvantages:

- Problems of reliability and validity (because the information is collected by a number of ***different persons*** who may have used ***different definitions*** or methods of obtaining data).

- There is a possibility that errors may occur when the information is extracted from the records. (This may be an important source of unreliability if ***handwritings*** are difficult to read.

- Since the records are maintained ***not for research*** purposes, but for clinical, administrative or ***other ends,*** the information required may not be recorded at all, or only partly recorded.

Interviews and self-administered questionnaires

Interviews may be less or more ***structured.*** A public health worker conducting interviews may be armed with ***a checklist of topics,*** but may not decide in advance precisely what questions he will ask. If his approach is ***flexible;*** the ***content, wording*** and ***order*** of his questions vary from interview to interview. Hence, his interviews are relatively ***unstructured.***

On the other hand, if a more standardized technique where the wording and order of the questions being decided in advance is used, it may take the form of ***a highly structured interview.***

Self-administered questionnaire: the respondent reads the questions and fills in the answers by himself (sometimes in the presence of an interviewer who “stands by” to give assistance if necessary.

The use of self-administered questionnaires is simpler and cheaper, such questionnaires can be administered to ***many persons simultaneously.***

Example:

- to students of a school
- they can also be sent by post unlike interviews.

However, they demand a certain **level of education** on the part of the respondent.

On the other hand, interviews have many advantages:

- A good interviewer can stimulate and maintain the respondents interest
  → the frank answering of questions.
- If anxiety is aroused (e.g., why am I being asked these questions?), the interviewer can allay it.
- An interviewer can repeat questions which are not understood, and give standardized explanations where necessary.
- An interviewer can ask “follow-up” or “probing” questions to clarify a response.
- An interviewer can make observations during the interview; i.e.,, note is taken not only of what the subject says but also how he says it.

In general, apart from their expense, interviews are preferable to self-administered questionnaires provided that they are conducted by ***skilled interviewers.***

While interviewing, a precaution should be taken not to influence the responses; the interviewer should ask his questions in a **neutral manner**. He should not show **agreement, disagreement**, or **surprise,** and should record the respondent’s precise answers without shifting or interpreting them.

Questionnaire Design

Questions may take two general forms: they may be “Open ended” questions, which the subject answers in his own words, or “closed” questions, which are answered by choosing from a number of fixed alternative responses.

In questionnaire design remember to:

a) Use familiar and appropriate language
b) Avoid abbreviations, double negatives, etc.
c) Avoid two elements to be collected through one question
d) Pre-code the responses to facilitate data processing
e) Avoid embarrassing and painful questions
f) Watch out for ambiguous wording
g) Avoid language that suggests a response
h) Start with simpler questions
i) Ask the same question to all respondents
j) Provide other, or don't know options where appropriate
k) Provide the unit of measurement for continuous variables (years, months, kgs, etc)
l) For open ended questions, provide sufficient space for the response
m) Arrange questions in logical sequence
n) Group questions by topic, and place a few sentences of transition between topics
o) Provide complete training for interviewers
p) Pretest the questionnaire on 20-50 respondents in actual field situation
q) Check all filled questionnaire at field level
r) Include "thank you" after the last question

Importance of combining different data-collection techniques

A skillful use of a combination of different data-collection techniques can maximize the quality of the data collected and reduce the chance of bias. Investigators often use a combination of flexible and less flexible research techniques.

Flexible techniques, such as, loosely structured interviews using open-ended questions and focus group discussions are called **qualitative research techniques**. They produce qualitative information, which is often recorded in narrative form.

Structured questionnaires that enable the researcher to quantify pre- or post-categorized answers to questions are an example of **quantitative research techniques**. The answers to questions can be counted and expressed numerically.

Both qualitative and quantitative research techniques are often used within a single study.

Methods of collecting qualitative data

Qualitative approaches to data collection usually involve direct interaction with individuals on a one to one basis or in a group setting. Data collection methods are time consuming and consequently data is collected from smaller numbers of people than would usually be the case in quantitative approaches such as the questionnaire survey. The benefits of using these approaches include richness of data and deeper insight into the phenomena under study.

Unlike quantitative data, raw qualitative data cannot be analysed statistically. The data from qualitative studies often derives from face-to-face interviews, focus groups or observation and so tends to be time consuming to collect. Samples are usually smaller than with quantitative studies and are often locally based. Data analysis is also time consuming and consequently expensive.

The main methods of collecting qualitative data are: individual interviews, focus groups and observation

Qualitative interviews

Qualitative interviews are semi structured or unstructured. If the interview schedule is too tightly structured this may not enable the phenomena under investigation to be explored in terms of either breadth or depth. Semi structured interviews tend to work well when the interviewer has already identified a number of aspects he wants to be sure of addressing. The interviewer can decide in advance what areas to cover but is open and receptive to unexpected information from the interviewee. This can be particularly important if a limited

time is available for each interview and the interviewer wants to be sure that the "key issues" will be covered.

Semi structured interviews (sometimes referred to as focused interviews) involve a series of open ended questions based on the topic areas the researcher wants to cover. The open ended nature of the question defines the topic under investigation but provides opportunities for both interviewer and interviewee to discuss some topics in more detail. If the interviewee has difficulty answering a question or provides only a brief response, the interviewer can use cues or prompts to encourage the interviewee to consider the question further. In a semi structured interview the interviewer also has the freedom to probe the interviewee to elaborate on the original response or to follow a line of inquiry introduced by the interviewee.
Unstructured interviews (referred to as "depth" or "in depth" interviews) have very little structure at all. The interviewer goes into the interview with the aim of discussing a limited number of topics, sometimes as few as one or two, and frames the questions on the basis of the interviewee's previous response. Although only one or two topics are discussed they are covered in great detail. Subsequent questions would depend on how the interviewee responded.

Unstructured interviews are exactly what they sound like - interviews where the interviewer wants to find out about a specific topic but has no structure or preconceived plan or expectation as to how they will deal with the topic. The difference with semi structured interviews is that in a semi structured interview the interviewer has a set of broad questions to ask and may also have some prompts to help the interviewee but the interviewer has the time and space to respond to the interviewees responses.

Qualitative interviews should be fairly informal. Interviewees should feel as though they are participating in a conversation or discussion rather than in a formal question and answer situation. However, achieving this informal style is dependent on careful planning and on skill in conducting the interview.

Focus group discussion

Sometimes it is preferable to collect information from groups of people rather than from a series of individuals. Focus groups can be useful to obtain certain types of information or when circumstances would make it difficult to collect information using other methods to data collection.

I. Characteristics and uses of focus group discussions

A focus group discussion (FGD) is a group discussion of 6-12 persons guided by a facilitator, during which group members talk freely and spontaneously about a certain topic.

The purpose of an FGD is to obtain in-depth information on concepts, perceptions, and ideas of the group. It aims to be more than a question-answer interaction.

FGD techniques can be used to:

a) Develop relevant research hypotheses by exploring in greater depth the problem to be investigated and its possible causes.
b) Formulate appropriate questions for more structured, larger scale surveys.
c) Supplement information on community knowledge, beliefs, attitudes, and behaviour already available but incomplete or unclear.
d) FGDs are not used to test hypotheses or to produce research findings that can be generalized.

II. Conducting a focus group discussion

Recruitment of participants: Participants should be roughly of the same socioeconomic group or have a similar background in relation to the issue under consideration. The age and sex composition of the group should facilitate free discussion.

If we need to obtain information on a topic from several different categories of informants who are likely to discuss the issue from different perspectives, we should organize a focus group for each category. For example, a group for men and a group for women.

Physical arrangements: Communication and interaction during the FGD should be encouraged in every possible way. Arrange the chairs in a circle. Make sure the area will be quiet, adequately lighted, etc., and that there will be no disturbances.

Preparation of a discussion guide: There should be a **written list** of topics to be covered. It can be formulated as a series of open-ended questions.

During the discussion: One of the members of the research team should act as a "**facilitator"** for the focus group. One should serve as "**recorder.**"

Functions of the facilitator:

- Introduce the session
- Encourage discussion
- Encourage involvement
- Listen carefully and move the discussion from topic to topic. Subtly control the time allocated to various topics so as to maintain interest.
- Take time at the end of the meeting to summarize, check for agreement and thank the participants.

In general, the facilitator should not act as an expert on the topic. His or her being there is to stimulate and support discussion.

Report writing in focus group discussions: Start with a description of the selection and composition of the groups of participants and a commentary on the group process, so the reader can assess the validity of the reported findings.

Present your findings, following your list of topics and guided by the objective(s) of your FGD. Include questions whenever possible, particularly for key statements.

The method of data collection chosen for a study should be appropriate for the type of information required. Whether the required information is quantitative or qualitative in nature is the major consideration. It would be time wasting to use unstructured interviews for

essentially quantitative studies where information could be more efficiently collected through structured interviews or questionnaires. Conversely, self completed questionnaires are generally unsuited to qualitative research: even when there is space for comments or for respondents to express ideas the space is limited and requires respondents to have skills in articulation and literacy.

Observation

Not all qualitative data collection approaches require direct interaction with people. It is a technique that can be used when data collected through other means can be of limited value or is difficult to validate. For example, in interviews participants may be asked about how they behave in certain situations but there is no guarantee that they actually do what they say they do. Observing them in those situations is more reliable: it is possible to see how they actually behave. Observation can also serve as a technique for verifying or nullifying information provided in face to face encounters.

In some research observation of people is not required but observation of the environment. This can provide valuable background information about the environment where a research project is being undertaken. For example, an action research project involving an institution may be enhanced by some description of the physical features of the building.

Bias in Information Collection

BIAS in information collection is a distortion in the collected data so that it does not represent reality.

Possible sources of bias during data collection:

1. Defective instruments, such as:

- Questionnaires with:

  — fixed or closed questions on topics about which little is known (often asking the 'wrong things');

  — open-ended questions without guidelines on how to ask (or to answer) them;

— vaguely phrased questions;

— ‘leading questions’ that cause the respondent to believe one answer would be preferred over another; or

— questions placed in an illogical order.

- Weighing scales or other measuring equipment that are not standardised.

These sources of bias can be prevented by **carefully planning the data collection process** and by **pre-testing the data collection tools.**

2. Observer bias:

Observer bias can easily occur when conducting observations or utilising loosely structured group- or individual interviews. There is a risk that the data collector will only see or hear things in which (s)he is interested or will miss information that is critical to the research.

Observation protocols and **guidelines for conducting loosely structured interviews** should be prepared, and training and practice should be provided to data collectors in using both these tools. Moreover it is highly recommended that data collectors **work in pairs** when using flexible research techniques and discuss and interpret the data immediately after collecting it. Another possibility - commonly used by anthropologists - is using a tape recorder and transcribing the tape word by word.

3. Effect of the interview on the informant:

This is a possible factor in all interview situations. The informant may mistrust the intention of the interview and dodge certain questions or give misleading answers**. For example:** in a survey on alcoholism you ask school children: ‘Does your father sometimes get drunk?’ Many will probably deny that he does, even if it is true. Such bias can be reduced by adequately introducing the purpose of the study to informants, by phrasing questions on sensitive issues in

a positive way, by taking sufficient time for the interview, and by assuring informants that the data collected will be confidential.

It is also important to be careful in the selection of interviewers. In a study soliciting the reasons for the low utilisation of local health services, for example, one should not ask health workers from the health centres concerned to interview the population. Their use as interviewers would certainly influence the results of the study.

4. Information bias:

Sometimes the information itself has weaknesses. Medical records may have many blanks or be unreadable. This tells something about the quality of the data and has to be recorded. For example, in a TB defaulter study the percentage of defaulters with an incomplete or missing address should be calculated.

Another common information bias is due to gaps in people's memory; this is called *memory* or *recall bias.* A mother may not remember all details of her child's last diarrhoea episode and of the treatment she gave two or three months afterwards. For such common diseases it is advisable to limit the period of recall, asking, for example, 'Has your child had diarrhoea over the past two weeks?'

Data quality checks at the time of the interview:

There are several ways of checking the quality of data. Some are given below:

- To examine consistency of data, two or more questions are asked at the beginning and these same questions are asked at the end. This helps to find out their consistency in the response.
- Repeat the question in another form to avoid doubts
- Use strict supervision and checking on the spot and re-interviewing
- For data completeness, check right at the end of the interview
- Check for odd answers given

Confidentiality of information

- Should be stated right on the top of the first page of the questionnaire
- If possible, use code numbers instead of names
- The purpose of the study should be explained at the beginning
- The respondent has the right not to be interviewed

6.10 Plan for data processing and analysis

Data processing and analysis should start in the field, with checking for completeness of the data and performing quality control checks, while sorting the data by instrument used and by group of informants. Data of small samples may even be processed and analyzed as soon as it is collected.

Why is it necessary to prepare a plan for processing and analysis of data?

Such a plan helps the researcher assure that at the end of the study:

- all the information (s)he needs has indeed been collected, and in a standardized way;
- (s)he has not collected unnecessary data which will never be analyzed.

The plan for data processing and analysis must be made after careful consideration of the objectives of the study as well as of the tools developed to meet the objectives. The procedures for the analysis of data collected through qualitative and quantitative techniques are quite different.

- For **quantitative data** the starting point in analysis is usually a description of the data *for each variable* for all the study units included in the sample. Processing of data may take place during data collection or when all data has been collected; description and analysis are usually carried out *after* the fieldwork has been completed.
- For **qualitative data** it is more a matter of describing, summarizing and interpreting the data obtained *for each study unit* (or for each group of study units). Here the researcher starts analyzing *while* collecting the data so that questions that remain

unanswered (or new questions which come up) can be addressed before data collection is over.

Preparation of a plan for data processing and analysis will provide you with better insight into the feasibility of the analysis to be performed as well as the resources that are required. It also provides an *important review of the appropriateness of the data collection tools* for collecting the data you need. That is why you have to plan for data analysis *before* the pre-test. When you process and analyze the data you collect during the pre-test you will spot gaps and overlaps which require changes in the data collection tools before it is too late!

What should the plan include?

When making a plan for data processing and analysis the following issues should be considered:

- Sorting data,
- Performing quality-control checks,
- Data processing, and
- Data analysis.

Sorting data

An appropriate system for sorting the data is important for facilitating subsequent processing and analysis.

If you have different study populations (for example village health workers, village health committees and the general population), you obviously would number the questionnaires *separately*.

In a comparative study it is best to sort the data right after collection into the two or three groups that you will be comparing during data analysis.

> **For example**, in a study concerning the reasons for low acceptance of family planning services, users and non-users would be basic categories; in a case-control study obviously the cases are to be compared with the controls.

It is useful to number the questionnaires belonging to each of these categories *separately* right after they are sorted.

> **For example**, the questionnaires administered to users of family planning services could be numbered U1, U2, U3, etc., and those for the non-users N1, N2, N3, etc.

Performing quality control checks

Usually the data have already been checked in the field to ensure that all the information has been properly collected and recorded. Before and during data processing, however, the information should be checked again for completeness and internal consistency.

If a questionnaire has not been filled in completely you will have MISSING DATA for some of your variables. If there are many missing data in a particular questionnaire, you may decide to exclude the whole questionnaire from further analysis.

- If an inconsistency is clearly due to a mistake made by the researcher/research assistant (for example if a person in an earlier question is recorded as being a non-smoker, whereas all other questions reveal that he is smoking), it may still be possible to check with the person who conducted the interview and to correct the answer.
- If the inconsistency is less clearly a mistake in recording, it may be possible (in a small scale study) to return to the respondent and ask for clarification.
- If it is not possible to correct information that is clearly inconsistent, you may consider excluding this particular part of the data from further processing and analysis as it will affect the validity of the study. If a certain question produces ambiguous or vague answers throughout, the whole question should be excluded from further analysis. (Normally, however, you would discover such a problem during the pre-test and change the wording of the question.)

For computer data analysis, quality control checks of data must also include a verification of how the data has been transformed into codes and subsequently entered into the computer. The same applies if data are entered into master sheets.

Data processing – quantitative data

Decide whether to process and analyse the data from questionnaires:

- **manually**, using data master sheets or manual compilation of the questionnaires, or
- **by computer**, for example, using a micro-computer and existing software or self-written programmes for data analysis.

Data processing in both cases involves:

- categorising the data,
- coding, and
- summarising the data in data master sheets, manual compilation without master sheets, or data entry and verification by computer.

1. Categorising

Decisions have to be made concerning how to categorise responses.

For **categorical variables** that are investigated through closed questions or observation, the categories have been decided upon beforehand.

In interviews the answers to open-ended questions (for example**,** 'Why do you visit the health centre?') can be pre-categorised to a certain extent, depending on the knowledge of possible answers that may be given. However, there should always be a category called 'Others, specify . . .', which can only be categorised afterwards.

These responses should be listed and placed in categories that are a logical continuation of the categories you already have. Answers that are difficult or impossible to categorise may be put in a separate residual category called 'others', but this category should not contain more than 5% of the answers obtained.

For **numerical variables,** the data are often better collected without any pre-categorisation. If you do not exactly know the range and the dispersion of the different values of these variables when you collect your sample (e.g., home-clinic distance for out-patients, or income), decisions concerning how to categorise and code the data at the time you develop your tools may be premature. If you notice during data analysis that your categories had been wrongly chosen you cannot reclassify the data anymore.

2. Coding

If the data will be entered in a computer for subsequent processing and analysis, it is essential to develop a CODING SYSTEM.

For computer analysis, each category of a variable can be coded with a letter, group of letters or word, or be given a number. For example, the answer 'yes' may be coded as 'Y' or 1; 'no' as 'N' or 2 and 'no response' or 'unknown' as 'U' or 9.

The codes should be entered on the questionnaires (or checklists) themselves. When finalising your questionnaire, for each question you should insert a box for the code in the right margin of the page. These boxes should not be used by the interviewer. They are only filled in afterwards during data processing. Take care that you have as many boxes as the number of digits in each code.

Coding conventions

Common responses should have the same code in each question, as this minimises mistakes by coders.

For example:

Yes (or positive response)	code - Y or 1
No (or negative response)	code - N or 2
Don't know	code - D or 8
No response/unknown	code - U or 9

Codes for open-ended questions (in questionnaires) can be done only after examining a sample of (say 20) questionnaires. You may group similar types of responses into single categories, so as to limit their number to at most 6 or 7. If there are too many categories it is difficult to analyze the data.

2. Summarizing the data in data master sheets, manual compilation, or compilation by computer

(1) Data master sheets

If data are processed by hand, it is often most efficient to summaries the raw research data in a so-called **DATA MASTER SHEET**, to facilitate data analysis. On a data master sheet all the answers of individual respondents are entered by hand.

To illustrate the use of master sheets, we will give an example of a rapid appraisal carried out by students of a nursing school about the smoking habits of the inhabitants of their town. The questionnaire had only 17 questions, of which 9 were asked of everyone, 4 exclusively to smokers and 4 exclusively to non-smokers. It was therefore decided to process the data by hand, divided in two groups: smokers and non-smokers, which were again subdivided in males and females. For each of the four groups, master sheets were prepared, on which all the answers of individual respondents could be recorded.

Master sheets can be made in different ways. For short simple questionnaires you may put all possible answers for each question in headers at the top of the sheet and then list or tick the answers of the informants one by one in the appropriate columns.

For example, the straightforward answers of the smoking questionnaire for male smokers could be processed as follows: **Master sheet for smokers (males)**

No.	Q1 Sex	Q2 Age		Q6 No of cig.		Q7 Age onset		Q9 Tried to				Q14 Cough > 2wks		Q14 Cough/ chest ever	
								reduce		stop					
		Yrs	Cat	No	Cat	Yrs	Cat	Yes	No	Yes	No	Yes	No	Yes	No
1	M	18	(1)	10	(2)	12	(2)	1x			√		√		√
2	M	35	(3)	30	(4)	20	(4)		NR	1x			√	√	
3	M	54	(4)	15	(2)	14	(2)	10x		3x		√		√	
Etc.															
Total	31	Av	35	Av	20	Av	18	26	4 + 1NR	19	12	5	26	11	20

Categories

Age	No. of cigarettes/day	Age onset smoking
15 - 24 = 1	< 10 = 1	< 10 = 1
25 - 34 = 2	10 - 19 = 2	10 - 14 = 2
35 - 44 = 3	20 - 29 = 3	15 - 19 = 3
45 - 54 = 4	30 - 39 = 4	20 - 24 = 4
55 + = 5	40 + = 5	25 + = 5

Note that for age and number of cigarettes smoked both the raw data and the categories have been entered. This makes it easier to control for coding mistakes and allows for calculating averages. There are 31 male smokers; if there are less than 31 answers, there must be some non-responders (NR), as happened in Q9, or a mistake was made. If you work

with *two persons,* one reading and one writing, the risk of mistakes will be reduced, as you can discuss the answers and control for mistakes while filling in the data.

(2) Compilation by hand (without using master sheets)

When the sample is small (say less than 30) and the collected data is limited, it might be more efficient to do the compilation manually.

Certain procedures will help ensure accuracy and speed.

1. If only one person is doing the compilation use **manual sorting**. If a team of 2 persons work together use either manual sorting or **tally counting**.
2. To do **manual sorting** the basic procedure is to:
    - Take one question at a time, for example, 'use of health facility',
    - Sort the questionnaires into different piles representing the various responses to the question, e.g., hospital/ health centre/ traditional practitioners) and
    - Count the number in each pile.

    When you need to sort out subjects who have a certain combination of variables (e.g. females who used each type of health facility) sort the questionnaires into piles according to the first question (gender), then subdivide the piles according to the response to the other question (use of health facility).
3. To do **tally counting** the basic procedure is:
    - One member of the compiling team reads out the information while the other records it in the form of a tally (e.g., /// representing 3 subjects, //// representing 4 subjects who present a particular answer).
    - Tally count for no more than two variables at one time (e.g., sex plus type of facility used).

    If it is necessary to obtain information on 3 variables (e.g., sex by time of attending a health facility by diagnosis), do a manual sorting for the first question, then tally count for the other two variables.

- After tally counting, add the tallies and record the number of subjects in each group.

4. After doing either manual or tally counting, **check** the total number of subjects/responses in each question to make sure that there has been no omission or double count.

Note:

One can tally in two ways, 𝍸 or 𝍷𝍸. The latter way is preferable as it reduces the possibility of error.

It should be noted that hand tallying is often used in combination with master sheet analysis when the relationship between two or three variables needs to be established, or details analyzed.

(3) Computer compilation

Before you decide to use a computer, you have to be sure that it will save time or that the quality of the analysis will benefit from it. Note that feeding data into a computer costs time and money. The computer should not be used if your sample is small and the data is mainly generated by open questions (qualitative data), unless there is a resource person who is competent in using a program for qualitative data analysis. The larger the sample, the more beneficial in general the use of a computer will be.

Computer compilation consists of the following steps:

1. Choosing an appropriate computer program
2. Data entry
3. Verification or validation of the data
4. Programming (if necessary)
5. Computer outputs/prints

i. Choosing an appropriate computer program

The identification of an appropriate statistical package is the first step in using a computer. Some examples of packages commonly used are: Epi Info, SPSS, STATA, etc.

ii. Data entry

To enter data into the computer you have to develop a data entry format, depending on the program you are using. After deciding on a data entry format, the information on the data collection instrument will have to be coded (e.g., Male: M or 1, Female: F or 2). During data entry, the information relating to each subject in the study is keyed into the computer in the form of the relevant code.

iii. Verification

During data entry, mistakes will definitely creep in. The computer can print out the data exactly as it has been entered, so the printout can be checked visually for obvious errors, (e.g., exceptionally long or short lines, blanks that should not be there, alphabetic codes where numbers are expected, obviously wrong codes).

Example:

- Codes 3-8 in the column for sex where only 1(F) and 2 (M) are possible
- Codes above 250 when you had only 250 subjects

If possible, computer verification should be built in. This involves giving the appropriate commands to identify errors.

Example:

The computer can be instructed to identify and print out all subjects where the 'sex' column has a code different from 1 (F) or 2 (M).

iv. Programming

A certain amount of basic knowledge of computer programming is needed to give the appropriate commands.

v. Computer outputs

The computer can do most of the analysis and the results can be printed. It is important to decide whether each of the tables, graphs, and statistical tests that can be produced makes sense and should be used in your report. That is why we PLAN the data analysis BEFOREHAND!

Data analysis – quantitative data

Analysis of quantitative data involves the production and interpretation of frequencies, tables, graphs, etc., that describe the data.

1. Frequency counts

From the data master sheets, simple tables can be made with **frequency counts** for each variable. A frequency count is an enumeration of how often a certain measurement or a certain answer to a specific question occurs.

For example,

Smokers	51
Non-smokers	93
Total	144

If numbers are large enough it is better to calculate the frequency distribution in percentages **(relative frequencies)**: 51/144 x 100 = 35% are smokers and 93/144 x 100 = 65% non-smokers. This makes it easier to compare groups than when only absolute numbers are given. In other words, percentages standardise the data.

It is usually necessary to summarise the data from numerical variables by dividing them into categories. This process may include the following steps:

(1) Inspect all the figures: What is their range? (The range is the difference between the largest and the smallest measurement.)

(2) Divide the range into three to five categories. You can either aim at having a reasonable number in each category (e.g. 0-2 km, 3-4 km, 5-9 km, 10+ km for home-clinic distance) or you can define the categories in such a way that they are each equal in size (e.g., 20-29 years, 30-39 years, 40-49 years, etc.). Sometimes one looks actively for a 'critical' value, when making different categories. For example, in a study relating family income to prevalence of diarrhoea over a certain period, there appeared to be no statistical relation when income was arbitrarily subdivided into four categories. When the *average* income was calculated, however, this

appeared to be a critical value. The children in families with an income above average had had significantly less diarrhoea than the children in families with an income below average.

(3) Construct a table indicating how data are grouped and count the number of observations in each group.

2. Cross-tabulations

Further analysis of the data usually requires the combination of information on two or more variables in order to describe the problem or to arrive at possible explanations for it.

For this purpose it is necessary to design CROSS-TABULATIONS.

Depending on the objectives and the type of study, two major kinds of cross-tabulations may be required:

- Descriptive cross-tabulations that aim at describing the problem under study.
- Analytic cross-tabulations in which groups are compared in order to determine differences, or which focus on exploring relationships between variables.

When the plan for data analysis is being developed, the data, of course, are not yet available. However, in order to visualize how the data can be organized and summarized it is useful at this stage to construct so-called DUMMY cross-tabulations.

A **DUMMY TABLE** contains all elements of a real table, except that the cells are still empty.

In a research proposal dummy tables should be prepared to describe the study population in order to show the crucial relationships between variables.

Some practical hints when constructing tables:

- If a dependent and an independent variable are cross-tabulated, the headings of the dependent variable are usually placed horizontally, and the headings of the independent variable vertically.
- All tables should have a clear title and clear headings for all rows and columns.
- All tables should have a separate row and a separate column for totals to enable you to check if your totals are the same for all variables and to make further analysis easier.

- All tables related to a certain objective should be numbered and kept together so the work can be easily organised and the writing of the final report will be simplified.

To further analyse and interpret the data, certain calculations or **statistical procedures** must usually be completed. Especially in large cross-sectional surveys and in comparative studies, statistical procedures are necessary if the data are to be adequately interpreted. Statistical tests should, for example, indicate whether differences are true differences or due to chance. When conducting such studies it is advisable to consult a person with statistical knowledge from the start in order that:

- correct sampling methods are used and an appropriate sample size is selected;
- decisions on coding are made that will facilitate data processing and analysis; and
- a clear understanding is reached concerning plans for data processing, analysis and interpretation, including agreement concerning which variables need to be cross-tabulated.

Processing and analysis of qualitative data

Qualitative data may be collected through open-ended questions in self-administered questionnaires, in individual interviews or focus group discussions or through observations during fieldwork. We will concentrate here on the analysis of responses obtained from open-ended questions in interviews or self-administered questionnaires.

Commonly solicited data in open-ended questions include:

- opinions of respondents on a certain issue;
- reasons for a certain behaviour; and
- descriptions of certain procedures, practices or perceptions with which the researcher is not familiar.

The data can be analyzed in seven steps:

Step 1: Take a sample of (say 20) questionnaires and list all answers for a particular question. Take care to include the source of each answer you list (in the case of questionnaires you can use the questionnaire number), so that you can place each answer in its original context, if required.

Step 2: To establish your categories, you first read carefully through the whole list of answers. Then you start giving codes (A, B, C, for example or **key words**) for the answers that you think belong together in one category, and write these codes in the left margin. Use a pencil so that it is easy to change the categories if you change your mind.

Step 3: List the answers again, grouping those with the same code together.

Step 4: Then interpret each category of answers and try to give it a label that covers the content of all answers. In the case of data on **opinions**, for example, there may be only a limited number of possibilities, which may range from (very) positive, neutral, to (very) negative.

Data on **reasons** may require different categories depending on the topic and the purpose of your question. After some shuffling you usually end up with 5 to 7 categories.

Step 5: Now try a next batch of 20 questionnaires and check if the labels work. Adjust the categories and labels, if necessary.

Step 6: Make a final list of labels for each category and give each label a code (keyword, letter or number).

Step 7: Code all your data, including what you have already coded, and enter these codes in your master sheet or in the computer.

Note again that you may include a category ‘others’, but that it should be as small as possible, preferably used for less than 5% of the total answers.

If you categories your responses to open-ended questions in this way you can:

- Analyze the content of each answer given in particular categories, for example, in order to plan what actions should be taken (e.g., for health education). Gaining *insight* in a problem, or in possible interventions for a problem, is the most important function of qualitative data.
- Report the number and percentage of respondents that fall into each category; so that you gain insight in the relative weight of different opinions or reasons.

Questions that ask for descriptions of procedures, practices, or beliefs usually do not provide quantifiable answers (though you may quantify certain aspects of them). The answers rather

form part of a jigsaw puzzle that you have to put together in order to obtain insight in your problem/topic under study.

IN CONCLUSION, a plan for the processing and analysis of data may include:

- a decision on whether all or some parts of the data should be *processed by hand or computer*;
- *dummy tables* for the description of the problem, the comparison of groups (if applicable) and/or the establishment of relationships between variables, guided by the objectives of the study;
- a decision on the *sequence* in which tables or data from different study populations should be analysed;
- a decision on how *qualitative data* should be analysed;
- an estimate of the *total time needed* for analysis and how long particular parts of the analysis will take;
- a decision concerning whether *additional staff* will be required for the analysis; and
- an estimate of the *total cost* of the analysis.

6.11 Ethical considerations

Why do we need ethical approval?

Before you embark on research with human subjects, you are likely to require ethical approval. You may wonder why all this bureaucracy is needed. But history shows us that prior to the development of ethical and human rights over the last 40 years, patients' rights were often ignored and many individuals were seriously harmed by medical experimentation.

- Atrocities committed during World War II in the Nazi Germany which led to the 1947 Nuremberg Code of Practice and in turn the 1964 Declaration of Helsinki

- Tuskegee Syphilis Study in USA (1932-1970s) to study the long-term effects of untreated syphilis- 400 men out of the 600 participants were never told about the infection and were never treated despite the fact that treatment became available
- A study to examine the natural progression of cervical carcinoma in New Zealand (1980s)-conventional treatment was withheld from women in trial and women were not asked for their consent

Ethical decisions are based on three main approaches: duty, rights and goal-based. The goal-based approach assumes that we should try to produce the greatest possible balance of value over disvalue. Discomfort to one individual may be justified by the consequences for the society as a whole. According to the duty-based approach, your duty as a researcher is founded on your own moral principles. As a researcher, you will have a duty to yourself and to the individual who is participating in the research. So even if the outcome of the proposed research is for a good cause, if it involves the researcher lying or deceiving his subjects in some way, then this would be regarded as unethical. In the rights-based approach, the rights of the individual are assumed to be all-important, thus a subject's right to refuse must be upheld whatever the consequences for the research.

Research studies should be judged ethically on three sets of criteria, namely: **ethical principles, ethical rules, and also scientific criteria**. The latter is often neglected but is important since if a study is poor or the sample size insufficient then the study is not capable of demonstrating anything and consequently could be regarded as unethical.

Ethical principles

Autonomy- ***we ought to respect the right to self-determination***

In research autonomy is protected by ensuring that any consent to participate in the study is informed or real. This means it is not enough to explain something about your project to a particular subject, it is the understanding and free choice whether or not to participate that is the key issue. There must be no coercion of any sort.

Non-Maleficence- ***we ought not to inflict evil or harm***

This principle states that we may not inflict harm on or expose people to unnecessary risk as a result of our research project. This is particularly important if our subjects may not be competent in some way, such as, the ability to give informed consent.

Beneficence – ***we ought to further others' legitimate interests***

This is the principle that obliges us to take positive steps to help others pursue their interests. These interests clearly have to be legitimate.

Justice-*we ought to ensure fair entitlement to resources*

This principle is concerned with people receiving their due. This means people should be treated equally in every way since not all people are equally competent or equally healthy.

Ethical rules

The ethical rules of research, like principles, are not absolute in that one may override another although clearly this must be justified. These rules are essential for the development of trust between researchers and study participants. Like the ethical principles on which the rules are based, there are four:

Veracity

All subjects in any research project should always be told the truth. There is no justification for lying, but this is not the same non-disclosure of information should it, in particular, invalidate the research.

Privacy

When subjects enroll in a research study, they grant access to themselves, but this is not unlimited access. Access is a broad term and generally includes viewing, touch or having information about them.

Confidentiality

Although someone may grant limited access to him or herself, they may not relinquish control over any information obtained. Certainly, no information obtained with the patient's or

subject's permission from their medical records should be disclosed to any third person without that individual's consent. This applies to conversations too.

Fidelity

Fidelity means keeping our promises and avoiding negligence with information. If we agree for example, to send a summary of our research findings to participants in a study we should do so.

Applying to ethics committee

Remember that the key questions that the Ethics Committee will be asking are:

- Is the research valid?
  -How important is the research question?
  -Can the question be answered?

- Is the welfare of the research subject under threat?
  -What will participating involve?
  -Are the risks necessary and acceptable?

- Is the dignity of the research subject upheld?
  -Will consent be sought?
  -Will confidentiality be respected?

6.12 Pretest or pilot study

Before the collection of data can be started, it is necessary to test the methods and to make various practical preparations. Pretests or pilot studies allow us to **identify potential problems** in the proposed study.

- A pretest usually refers to a small-scale trial of a particular research component.

- A pilot study is the process of carrying out a preliminary study, going through the ***entire research procedure*** with a ***small*** sample.

Pretesting is: 1. Simpler
2. Less time consuming
3. Less costly than conducting an entire pilot study.

Therefore, pretesting is recommended as an essential step in the development of the research projects. It is useful in examining the **practicability**, **reliability** and **suitability** of the method.

The comments of the respondents will help in **improving the sequence** and **layout** of the questionnaire. It is also important to know the **time** taken by the interview.

In a community study, cooperation can be enhanced by suitable **public relations** and **preparatory educational work** in the community. The best results are provided by **contacts with key individuals** and **organizations** in the community.

6.13 Exercises

1. Identify the most appropriate study design for the research proposal you are planning to develop.
2. Describe the various data collection techniques and state their uses and limitations.
3. State the differences between quantitative and qualitative research methods by giving appropriate examples.
4. A nutritionist wants to determine the prevalence of malnutrition among under 5 children in Amhara region. If a sample of 3000 children is required, what is the sampling technique he should use to select the required subjects. Write a short note on the procedures (steps) he should follow in selecting these subjects.
5. In a school there are about 1800 students and the investigator wants to determine the prevalence of a certain character (eg., KAP on HIV/AIDS) by taking 450 students. The following table gives the distribution of students by grade and number of sections.

Grade	Number of students	Number of sections
9	600	8
10	500	7
11	400	6
12	300	5
Total	1800	26

a) What type of sampling technique do you use? Why?

b) How do you select the subjects who will be included in your sample?

6. A multi-national clinical trial is proposed to investigate the value of a gradually increasing dose schedule of a beta blocker in the treatment of severe heart failure. The trial will be randomised, double-blind and placebo controlled. Each patient is to be followed for 2 years, and the main treatment comparison is for all cause mortality. Previous experience suggests a 2 year mortality rate of around 30%. The investigators propose that a one-third reduction in mortality due to beta-blockade would be important to detect. They suggest that type I and type II errors be set at .05 and .1, respectively.

   a) Calculate the required number of patients to be recruited.
   b) Suppose one anticipates that 10% of patients randomised to beta-blockade will fail to comply with the intended treatment policy. What change in required sample size would you suggest?

7. Prepare your data-collection tools, taking care that you cover all important variables of your proposed study.

CHAPTER SEVEN

WORK PLAN AND BUDGET

7.1 Work Plan

A WORK PLAN is a schedule, chart or graph that summarizes the different components of a research project and how they will be implemented in a coherent way within a specific time-span.

It may include:

- The tasks to be performed;
- When and where the tasks will be performed; and
- Who will perform the tasks and the time each person will spend on them.

Work plan could be presented in different forms, such as work schedule and GANTT chart, but we will demonstrate the GANTT chart here.

A GANTT chart is a planning tool that depicts graphically the order in which various tasks must be completed and the duration of each activity. The length of each task is shown by a bar that extends over the number of days, weeks or months the task is expected to take.

How can a work plan be used?

A work plan can serve as:

- A tool for planning the details of the project activities and drafting a budget.
- A visual outline or illustration of the sequence of project operations. It can facilitate presentations and negotiations concerning the project with government authorities and other funding agencies.
- A management tool for the Team Leader and members of the research team, showing what tasks and activities are planned, their timing, and when various staff members will be involved in various tasks.
- A tool for monitoring and evaluation, when the current status of the project is compared to what had been foreseen in the work plan.

Example of a GANTT Chart

		Responsibility	Month 1				Month 2				Month 3				Month 4				Month 5			
1.	Prepare proposal and submit to donors	PI																				
2.	Obtain fund and discuss arrangement with local government	PI																				
3.	Preparation of study tool	PI																				
4.	Prepare for field Work	PI``																				
5.	Travel to data collection site	PI																				
6.	Select data collectors and research assistants	PI																				
7.	Conducting training for data collectors and supervisors	PI																				
8.	Pre-testing of the survey instrument	PI+RA+DC																				
9.	Data collocation	PI+RA+DC																				
10.	Data entry and cleaning	PI+RA+DEC																				
11.	Data analysis and write up	PI																				
12.	Prepare workshop on findings	PI+RA																				
13.	Hold workshop	PI+RA																				

P.I. =Principal Investigator
R.A. = Research Assistant
D.C. = Data Collectors
DEC = Data entry clerk

7.2 Budget

Why do we need to design a budget?

- A detailed budget will help you to identify which resources are already locally available and which additional resources may be required.
- The process of budget design will encourage you to consider aspects of the work plan you have not thought about before and will serve as a useful reminder of activities planned, as your research gets underway.

How should a budget be prepared?

It is necessary to use the work plan as a starting point. Specify, for each activity in the work plan, what resources are required. Determine for each resource needed the **unit cost** and the **total cost**.

The budget for the fieldwork component of the work plan will include funds for personnel, transport and supplies.

Note that UNIT COST (e.g., per diem or cost of petrol per km), the MULTIPLYING FACTOR (number of days), and TOTAL COST are required for all budget categories.

Budget justification

It is not sufficient to present a budget without explanation. The budget justification follows the budget as an explanatory note justifying briefly, in the context of the proposal, why the various items in the budget are required. Make sure you give clear explanations concerning why items that may seem questionable or that are particularly costly are needed and discuss how complicated expenses have been calculated. If a strong budget justification has been prepared, it is less likely that essential items will be cut during proposal review.

Example of a budget proposal

	Budget Category	Unit Cost	Multiplying factor	Total Cost (Birr)
1.	Personnel	Daily Wage (including per diem)	Number of staff days (Number of staff x Number of working days)	
	Principal investigator	100	1x15	1,500.00
	Supervisors	100	2 x 15	3,000.00
	Data collectors	60	10 x 15	9,000.00
	Data entry clerk	40	1x 20	800.00
	Secretarial work	40	1x20	800.00
	Sub total		**Personnel TOTAL**	**15,100.00**
2.	Transport	Cost per km	Number of km	
			(no. vehicles x no. days x no. km)	
	Car	1 Birr	2 x 10 x 100 =2000	2,000.00
	Sub total		**Transport TOTAL**	**2,000.00**
3.	Supplies	Cost per Item	Number	
	Questionnaire duplication	3 Birr/Quest.	1500	4,500.00
	Clip board	16	13	208.00
	Flip chart paper	2	50	100
	Pen	1	30	30
	Pencil	0.25	30	7.50
	Eraser	0.50	30	15.00
	Sharper	0.50	30	15.00
	Marker	12	12	144.00
	Transparency (pack)	150	1	150.00
	Printing paper (pack)	40	4	160.00
	Photocopying cost	0.25	1000	250.00
	Printing and Binding	20	10	200.00
			Supplies TOTAL	**5,779.50**
4.	Training	Cost per item	Number of days	
	Hall rents	100	4 days	400
	Tea/coffee	10 Birr/participant/day (10x13)=130	4	520
			Training TOTAL	**920.00**
			Grand Total	**23,799.50**

CHAPTER EIGHT

MAJOR COMPONENTS AND OUTLINE OF THE DIFFERENT PHASES IN A RESEARCH PROCESS

8.1 Summary of the major components of a research proposal

The details of the development of a health research proposal (protocol) are given in the previous chapters. It is also important to give the summary of the major components and steps to assist students to have a general idea of the outline in a relatively short period of time.

It should be noted that the proposal will be designed after a topic is accepted to be researched. And, for approval, the protocol design is required to include at least the contents given below:

Title and cover page

The cover page should contain the title, the names of the authors with their titles and positions, the institution (e.g., Gondar College of Medicine and Health Sciences) and the month and year of submission of the proposal. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study will be undertaken.

Abstract: Summary of the proposal which should include (in short):
Objectives, hypothesis, methods, time schedule and the total cost.

Table of contents: A table of contents is essential. It provides the reader a quick overview of the major sections of your research proposal, with page references, so that (s)he can go through the proposal in a different order or skip certain sections

I) Introduction

- **Statement of the research problem**
  - Background and definition of the problem of the study

- Why the proposed study is important, i.e., general statement on rationale behind the research project.

- **State of knowledge: knowledge pertinent to subject under study**
  - Local data/knowledge
  - Literature review
- **Significance of the proposed work**

Specific statements on the significance of the results of the study should be given. Where to use the results; who to make us of the results; what for the result would be used; and other details related with the usefulness of the ***end results*** of the study.

II) Objective of the study

- General objective: aim of the study in general terms
- Specific objectives: measurable statements on the specific questions to be Answered
- Hypotheses

The objectives should meet the purpose of the study. They should be phrased clearly, unambiguously and very specifically. Also, they should be phrased in measurable terms.

III) Materials and methods

If the investigation deals with human beings, the terms 'study population' or 'subjects' are preferable to 'materials'.

- Type of study (study design)

- Study population
  - Describe the study areas and populations
  - Mapping and numbering of the study area
  - Appropriateness of the study
  - Accessibility (provide background information, travel, time, etc...)
  - Cooperation and stability of the population

- Type of data (defining each variable to be collected and methods for collecting them)

- Operational definitions
- Some elements of the variables to be studied:

What characteristics will be measured? How will the variables be defined? What scales of measurement will be used etc.

- Inclusion/ exclusion criteria
- Sampling procedure to be used and sample size and power calculation.
- Data collection and management
  - Data collection and coding forms should be appended to protocol
  - Training and quality control, bias control, data entry and storage, data clean-up and correction of deficiencies
- Data analysis
  - Management of dropouts
  - Frequencies, rates, other parameters
  - Statistical programs and tests to be used
  - Data presentation (dummy tables to be appended)
- Ethical considerations: rights and welfare of the subjects and method of obtaining their informed consent
- Pretest or pilot study:
  (allows us to identify potential problems in the proposed study)

IV) Work plan (project management)

- Personnel, job descriptions, training
- Schedule (timetable)- provide actual dates for each activity
  - Pilot phase
  - Final study
- Onset, data collection, analysis, write-up
- Relevant facilities
- Cooperating organizations

V) Budget (itemize all direct costs in Ethiopian Birr)

- Personnel, material/supplies, travel, analysis, contingency, etc.

VI) References: List only those cited in text and number by order they appear in text using Arabic numerals.

VII) Appendices: -

- Data collection and coding forms
- Dummy tables for data presentation
- Letters of support (cooperation)

Exercise

Develop a research proposal of your own topic. Take account of all the chapters covered so far and write your final proposal in line with the guideline given above.

8.2 Summary of the major activities of the fieldwork phase

Activities to be performed during the field period

1. **Briefing of managers and health service personnel:** The purpose of the briefing is to obtain support for the project. Such support is necessary to obtain resources as well as to get permission to collect the required data. Briefing should be conducted with all important persons and/or organizations at different levels.

2. **Identifying and obtaining project resources**: We have to identify and obtain the necessary resources (manpower, materials, etc.) needed to collect the required data. We have to make sure that all the items needed for the study are ready.

3. **Reviewing availability of subjects**: It is important to make personal visits to every site where the actual data will be collected to understand the physical and manpower limitations, constraints and special circumstances that could influence data collection. This would assist the investigators to take an appropriate measure and make the necessary preparations.

4. Organizing logistics for data collection: Having made an inventory of available resources, the logistics for data collection have to be organized. This will involve planning in detail how , where, and when data collection will be carried out.

5. Preparing fieldwork manuals: Manuals or instruction sheets should be prepared for interviewers (data collectors) and supervisors.

- The manual for data collectors should have instructions concerning the:
  - purpose of the study
  - role of the data collectors
  - the way data collectors should introduce themselves to respondents
  - interviewing techniques
  - questionnaire (general format, clarification of terms, instructions regarding how to ask sensitive or complicated questions, instructions concerning how to fill in answers, such as, answers to open-ended questions)
  - sampling procedures (and what to do if the required respondent is absent, etc.)

- The manual for supervisors, in addition to the above instructions, should include information on:
  - maintaining a record of data collectors' attendance
  - safe-keeping of completed (filled) data
  - determining the number of interviews to be completed each day by every data collector depending on the specific circumstance the data collector is found (e.g., the houses to be covered could be scattered, etc.)
  - ensuring the quality control of fieldwork
  - dealing with non-responses and incomplete interviews
  - reporting progress to the coordinator at specific intervals

6. Training of data collectors and supervisors

Data collectors and supervisors must be given explicit training. Their training should be supported by practical exercises. They should be involved in the pretest. Following the

pretest, they could participate in the adjustment of instruction sheets and data-collection tools.

During the training, the data collectors and supervisors should be strongly instructed that they would be responsible for any mistakes that may arise due to their negligence and lack of adherence to the manual.

7. Conducting pretest in the research location (nearby area) and revising data-collection tools

- The pretest should assess the validity of the data-collection instruments and procedures, as well as the sampling procedures
- It should identify scientific as well as logistical problems and constraints
- Revise the data-collection tools and other procedures after the pretest

8. Data collection

After getting permission for the conduct of the study, obtained the necessary resources, trained the required personnel, made pretests and modified data-collection tools and procedures, the required data collection can be carried out.

9. Processing data

After collecting and sorting the data, all questionnaires and records should be checked for errors. This should be done before leaving the area where data collection is done. If there is any error to be corrected regarding a particular questionnaire, it would be easy for the data collector or supervisor to make the changes by visiting the respondent from which the information was obtained.

During this stage, answers given for open-ended questions may be converted into quantifiable numerical form for processing by computer or other means.

8.3 Writing a research report

Writing a good report may take much time and effort. The most difficult task is usually the preparation of the first draft. The report should be easily intelligible. This requires clarity of language, a logical presentation of facts and inferences, the use of easily understood tables and charts, and an orderly arrangement of the report as a whole. It should be no longer than is necessary.

Conventionally, a report usually contains the following major components.

Title and cover page

The cover page should contain the title, the names of the authors with their titles and positions, the institution that is publishing the report, (e.g., Gondar College of Medicine and Health Sciences) and the month and year of publication. The title could consist of a challenging statement or question, followed by an informative subtitle covering the content of the study and indicating the area where the study was implemented.

Abstract (Summary)

The summary should be brief and informative. A reader who has been attracted by the title will usually look at the summary to decide whether the report is worth reading. The summary should be written only *after* the first or even the second draft of the report has been completed. It should contain:

— a very brief description of the problem (WHY this study was needed)

— the main objectives (WHAT has been studied)
— the place of study (WHERE)
— the type of study and methods used (HOW)
— major findings and conclusions, followed by
— the major (or all) recommendations.

The summary will be the first (and for busy health decision makers most likely the only) part of your study that will be read. Therefore, its writing demands thorough reflection and is time

consuming. Several drafts may have to be made, each discussed by the research team as a whole

Acknowledgements

It is good practice to thank those who supported you technically or financially in the design and implementation of your study. Also your employer who has allowed you to invest time in the study and the respondents may be acknowledged. Acknowledgements are usually placed right after the title page or at the end of the report, before the references.

Table of contents

A table of contents is essential. It provides the reader a quick overview of the major sections of your report, with page references, so that (s)he can go through the report in a different order or skip certain sections.

List of tables, figures

If you have many tables or figures it is helpful to list these also, in a 'table of contents' type of format with page numbers.

List of abbreviations (optional)

If abbreviations or acronyms are used in the report, these should be stated in full in the text the first time they are mentioned. If there are many, they should be listed in alphabetical order as well. The list can be placed before the first chapter of the report.

The table of contents and lists of tables, figures, abbreviations should be prepared last, as only then can you include the page numbers of all chapters and sub-sections in the table of contents. Then you can also finalise the numbering of figures and tables and include all abbreviations.

I) Introduction

The introduction is a relatively easy part of the report that can best be written after a first draft of the findings has been made. It should certainly contain some relevant (environmental/ administrative/ economic/ social) background data about the country, the health status of the

population, and health service data which are related to the problem that has been studied. You may *slightly comprise or make additions to the corresponding section in your research proposal,* including additional literature, and use it for your report.

Then the statement of the problem should follow, again revised from your research proposal with additional comments and relevant literature collected during the implementation of the study. It should contain a paragraph on what you hope(d) to achieve with the results of the study.

Global literature can be reviewed in the introduction to the statement of the problem if you have selected a problem of global interest. Otherwise, relevant literature from individual countries may follow as a separate literature review after the statement of the problem. You can also introduce theoretical concepts or models that you have used in the analysis of your data in a separate section after the statement of the problem.

II) Objectives

The general and specific objectives should be included as stated in the proposal. If necessary, you can adjust them slightly for style and sequence. However, you should not change their basic nature. If you have not been able to meet some of the objectives this should be stated in the methodology section and in the discussion of the findings. The objectives form the HEART of your study. They determined the methodology you chose and will determine how you structure the reporting of your findings.

III) Methods

The methodology you followed for the collection of your data should be described in detail. The methodology section should include a description of:

- the study type;
- major study themes or variables (a more detailed list of variables on which data were collected may be annexed);
- the study population(s), sampling method(s) and the size of the sample(s);
- data-collection techniques used for the different study populations;

- how the data were collected and by whom;
- procedures used for data analysis, including statistical tests (if applicable).

If you have deviated from the original study design presented in your research proposal, you should explain to what extent you did so and why. The consequences of this deviation for meeting certain objectives of your study should be indicated. If the quality of some of the data is weak, resulting in possible biases, this should be described as well under the heading 'limitations of the study'.

IV) Results

- Findings should be presented
- Tables and graphs could be used (should be well titled and captioned)
- The tables should be well constructed, and without anomalies such as percentages which do not add up to 100 percent
- Avoid too many decimal places
- Graphs should clarify and not complicate, and care should be taken that they do not mislead
- If appropriate statistical tests are used, the results should be included. P-values alone are not very helpful. Confidence intervals and the type of tests used should be indicated.

V) Discussion

The findings can now be discussed by objective or by cluster of related variables or themes, which should lead to conclusions and possible recommendations. The author interprets the findings. Care should be taken not to introduce new findings, i.e., findings not mentioned in the result section. The discussion may include findings from other related studies that support or contradict your own. Limitation of the study and generalizability of the finding should also be mentioned.

VI) Conclusions and recommendations

The conclusions and recommendations should follow logically from the discussion of the findings. Conclusions can be short, as they have already been elaborately discussed in chapter 5. As the discussion will follow the sequence in which the findings have been

presented (which in turn depends on your objectives) the conclusions should logically follow the same order.

It makes easy reading for an outsider if the recommendations are again placed in roughly the same sequence as the conclusions. However, the recommendations may at the same time be summarised according to the groups towards which they are directed, for example:

- policy-makers,
- health and health-related managers at district or lower level,
- health and health-related staff who could implement the activities,
- potential clients, and
- the community at large.

Remember that action-oriented groups are most interested in this section.

In making recommendations, use not only the findings of your study, but also supportive information from other sources. The recommendations should take into consideration the local characteristics of the health system, constraints, feasibility and usefulness of the proposed solutions. They should be discussed with all concerned before they are finalised.

VII) References

The references in your text can be numbered in the sequence in which they appear in the report and then listed in this order in the list of references (**Vancouver system**). Another possibility is the **Harvard system** of listing in brackets the author's name(s) in the text followed by the date of the publication and page number, for example: (Shan 2000: 84). In the list of references, the publications are then arranged in alphabetical order by the principal author's last name. You can choose either system as long as you use it consistently throughout the report.

VIII) Annexes or appendices

The annexes should contain any additional information needed to enable professionals to follow your research procedures and data analysis.

Information that would be useful to special categories of readers but is not of interest to the average reader can be included in annexes as well.

Examples of information that can be presented in annexes are:

- tables referred to in the text but not included in order to keep the report short;
- lists of hospitals, districts, villages etc. that participated in the study;
- questionnaires or checklists used for data collection.

CHAPTER NINE

DEFINITION OF COMMON TERMS APPLIED IN COMPUTER USE AND APPLICATION OF SOME STATISTICAL PACKAGES

9.1 Introduction to a microcomputer

A computer is a general purpose, electronic, stored-program device. It is designed to handle program instructions as well as the data to be processed by the instructions.

First generation computers (1950s) were:

- ***Large in size***
- ***Poor capacity***

Today, great advances have been made in computer technology

- ***Size improved (reduced)***
- ***High capacity***

A microcomputer is a small computer, but it is a powerful tool. It can help you to write letters and reports, to produce health books, to do work with health statistics, etc.

In the computer world, the term “user friendly” is often heard. This refers to the ease with which an individual can interact with the computer and understand what the computer has done for the user. The greater the user’s understanding of and familiarity with the computer, the friendlier the computer gets.

Two important words to learn are hardware and software. Hardware refers to the parts of the computer that you can see and touch. Software is a collection of programs that support the operations of a computer. It shows the parts that you cannot see or touch - it is the instructions given to the computer in its own language that tells it what to do and how to do it.

What is the hardware that you can see?

When you look at a computer you see many different pieces of equipment. What are their names and what do they do?

1. **VDU (visual display unit):** it works like a television because it allows you to see the work that you do with the computer.

2. **Motherboard:** it is a type of a metal box. The CPU (Central Processing Unit) is inside this box. It is the brain of the computer.

3. **Keyboard:** the keyboard looks a bit like a typewriter. The keys in the middle of the keyboard work like a typewriter and it is called the typewriter area of keyboard. On the right side of the keyboard there is a numeric pad. It has numbers on the keys like a pocket calculator. On the left and top of the keyboard are the function keys. These keys are used to give special instructions to the computer.

4. **Printer:** A printer is used to make paper copies of information that the computer has stored.

5. **Uninterruptable power supply (UPS):-** it takes electricity from the wall socket and sends it to the computer. It provides the steady source of electricity that the computer needs. The purpose of the UPS is to protect your computer work in the event of a power failure.

Basic components of a typical computer system

The basic components of a microcomputer system usually include four parts: the input unit, the central processing unit (CPU), auxiliary storage (i.e., secondary storage) and the output unit

1. **The input unit:** although there are different ways to enter information to a computer, the keyboard is the most important input device. One important point to remember about using the keyboard is that when you enter information, the information appears on the monitor but goes no farther. The information is understood by the computer only after the ***Enter key*** has been pressed. You must therefore press the Enter key at the end of every command line.

2. **The central processing unit (CPU)**

   All computations, regulation of data flow, decision-making operation, etc. are carried out by this part of the computer.

   The CPU is an important component of the computer system that contains the control unit, the arithmetic-logic unit (ALU), and main memory (Primary storage).

   All these parts work together to electronically control the functions of the computer system.

3. **Auxiliary (secondary) storage devices:**

   To store large amount of information, such as large database, you need auxiliary storage devices to supplement the internal memory. Common external storage devices include floppy disks (diskettes) and hard disks.

4. **The output unit:** the output unit in a computer system displays the results of computations or data processing. Output can be displayed temporarily on a monitor, or permanent "hard copy" can be produced on a printer.

9.2 Introduction to some common software packages

Word-Processing:

Microsoft Word is a word-processing program which is used to create and amend files containing textual information. It is capable of handling a wide range of applications. It is used for the production of simple documents such as letters and it handles complex-tasks (e.g. production of substantial reports). Before the introduction of Microsoft Word there were other word-processing programs like Word star and Word perfect.

Database:

In general terms, a database is any ordered collection of information. Thus a telephone directory or the list of the university of Gondar students would constitute a database.

A typical data set will consist of a collection of cases (or records), each of which contains the values of a set of variables (or fields).

Although it would be possible to write computer programs to organize and analyse a file of data, in general **Database management systems (DBMSS)** are now used. These are ready-written, general purpose programs to help in the organization and analysis of data. Two examples are: **Epi-Info and SPSS.**

I. Introduction to Epi Info 6

- Epi Info is a multi-purpose computer program designed for epidemiological researchers.
- Within Epi-Info there are smaller programs designed to perform specific tasks.

→ **EPED** (questionnaire design)

→ **ENTER** (data entry)

→ **CHECK, VALIDATE** (data checking)

→ **ANAYSIS** (data analysis)

→ **STATCALC** (simple statistics)

→ **etc.**

Getting Started

- On different computer systems there will be different ways of starting up the EPi Info program.
- However, in most cases, it will be sufficient to issue the command EPI6 at the MSDOS prompt and press the *Enter* key.

In the Epi Info program the EPED module has to perform three separate functions. These are:

- Wordprocessing and Report writing
- Creating questionnaire for use with ENTER and CHECK programs
- Editing command files for use with the ANALYSIS module

Making a questionnaire

- A questionnaire is the template that guides Epi Info in making a data file. Once you have created a questionnaire, making a data file is an automatic process.

- An Epi Info questionnaire may have up to 500 lines. Headings and other text may appear anywhere. Places where you will enter data are called "Fields" or (in the analysis phase), "Variables".

- The first item is usually a heading.

- Now, let's start the questions with "Name" and a series of **underline characters** to indicate the entry field. The number of characters allowed in the entry is indicated by the length of the blank:

Name _____

- When numbers (instead of letters) are entered, we use the symbol " # ".

Example:

Age ##

Sex #

Monthinc ###

Saving the questionnaire

- When you feel that you have completed preparing the template (questionnaire), press the function key " F2 " and take the cursor to **" Save file to..".**

- Press the Enter (Return) key while the cursor is on **" Save file to..".**

- Write the file name with the extension "qes"

Helpful Tips

a) File names should be recognizable ones and should not exceed eight characters. A character could be a letter or a number.

b) The extension for an Epi Info questionnaire file name should always be "qes".

Example

birthwt.qes

trachoma.qes

Exercise :

Develop a questionnaire of your own data and prepare a template using the EPED program.

Entering Data using the ENTER program

- The **ENTER** program will create a data file from a questionnaire. That is, The **ENTER** program will create an Epi Info database < **.rec** > file using the questionnaire.
- Once the < **.rec** > file is created, the file may be loaded into ENTER for adding more records or editing those already entered.
- If the questionnaire is revised, **ENTER** can be instructed to revise the < .rec > file accordingly.
- In the **ENTER** module, there are five options to choose from. We will see the first three ones which are in common use.

    1) Enter or edit data

    2) Create new data file from **.Qes** file

    3) Revise structure of data file using revised **.Qes**

- Now, move the cursor to the **ENTER** module and press the Enter key.

- Write the file name on the space given below "**Data file < .Rec> :**"

- Write the numbers **1** or **2** or **3** (depending on the type of tasks to be performed) on the blank space next to "**Choose one:** __".

- If you choose number **2,** write the same file name on the blank space under "**New Questionnaire file <.Qes**>". Finally, press the Enter key while the cursor is blinking on "Y".

Exercise :

Enter your data using the template you have developed.

CHECK customize entry

- Often it is useful to have the computer check for errors during the data entry process and to skip over parts of the questionnaire if certain conditions are met.
- The **CHECK** program instructs **ENTER** to perform such operations automatically.
- **CHECK** makes a file with a name ending in **.CHK**.
- The **.CHK** file contains instructions for ENER to restrict the data entered in specified fields.
- When **ENTER** is run, it automatically looks for a file with the same name as the **.Rec** file but ending in **.CHK**.
- Using **CHK** is optional.
- Now, move the cursor to "**CHECK customize entry**" and press the Enter key.
- Write the name of your **.Rec** file on the space below "**Data file < .Rec**>:
- Let's see two cases:

  a) **F1/F2 - (Min/Max)**

  - Put the cursor on the blank spaces of any of the fields (variables)
  - Write the minimum number (e.g., the minimum number for women aged 15 to 49 years is 15)

- **Press F1**
- Write the maximum number (e.g., the maximum number for women aged 15 to 49 years is 49)

- **Press F2**
- Finally, the valid values will be indicated at the bottom of your questionnaire. In this case, the valid values will be 15 to 49.

- **Save data**

b) **F7 - Jump**

Appropriate jumps can be built in so that questions which are not applicable as a result of a previous response are jumped.

- put the cursor on the blank spaces of any of the fields (variables).
- Write any number different from the valid ones

- **Press F7**
- Put cursor in field to jump to an entry of the above number
- Press **F7 again** and ENTER will automatically jump over the fields in between.

- **Save data**

Exercise:

Once you create the data entry format, use the **CHECK** module to customize your entry for some of the data values.

Analyzing Data

- **ANALYSIS** produces lists, frequencies, tables (cross tabulations), statistics and graphs.

- Move the cursor to "**ANALYSIS of data**" and press the Enter key.
- Use the "**READ**" command to choose a dataset.

  Example: **Read birthwet.rec**

- Once your **.Rec** file is retrieved, you can perform the tasks explained above.

  Example

  freq age
  freq sex
  tables age sex
  List sex
  Update age sex etc.

- If you press the function key "**F2**", you will see the various "commands" that perform different tasks.
- If you press the function key "**F3**", you will see the list of "variables" that you have created using the EPED program.
- If you **press** the function key "**F5**", your output will go to the printer. **If you press it again,** the output will go to the screen.

DEFINE and RECODE - These commands are very important in facilitating the analysis of data.

- **DEFINE** allows creation of new variables for use in analysis.
- Variable names in the DEFINE statement must be 10 or fewer characters.
- Variable names do not begin with a number.

Example:

DEFINE age1 ## (a newly created variable)

This could be used in a RECODE statement to provide labels for a numerically coded age variable.

- The RECODE command is used to form several categories (groups).

Example:

RECODE **variable1** to **variable2** 15-19=1 20-24=2 25-29=3 30-34=4 35-39=5 40-44=6 45-49=7

- Variable1 refers to the old variable while variable2 refers to the newly created one.
- The variable "age" given earlier is an old one. Therefore, we can put it in place of variable1
- The variable "age1" shown above is a new one. Therefore, we can put it in place of variable2.

After successfully completing the recoding of variables, you can use other commands (e.g., freq, tables, list, etc.) to have a summary statistics based on the new variables.

Helpful tips

a) The update command will assist you to see all the data values you have entered. You can make any changes using this command.

b) Pressing F10 will take you one step back.

STATCALC calculator

- **STATCALC** does statistical analysis of data entered from the keyboard into tables on the screen.
- Facilities are provided to perform **2 by n tables**, to investigate **linear trends** and to **calculate sample sizes.**

Example:

A cross-sectional survey on knowledge and use of condom was carried out among commercial sex workers (CSWs) of three small towns in Northwest Ethiopia. The table below shows the experience of these women on the use of condom classified by some selected socio-demographic variables. [*G.D. Alene, Ethiop.J.Health DEV.2002;16(3):277-286]*

Variable	Ever used condoms	
	Yes	No
Age of CSWs (in year)		
15 - 24	55	77
25 - 34	36	71
35 - 49	8	58
Educational status of CSWs		
Can't read and write	46	131
Can read and write (informal schooling)	14	21
Elementary school	22	32
High School	17	22
Occupation of partners		
Farmer (subsistence)	76	182
Trader (Private business)	15	13
Others	8	11

Investigate the influence of the above selected variables on the use of condom among the CSWs of the study areas.

- Move the cursor to STATCALC calculator and press the Enter key.
- Enter the data values on the appropriate cells.
- Use the ordinary Chi square test (2 by n) first.
- For ordinal type of data, consider the Chi square for trend test.

Influence of some socio-demographic variables on the use of condom among commercial sex workers who had knowledge about condoms, Northwest Ethiopia, Dec. 1999.

Variable	Ever used condoms (n = 305)		Odds Ratio	P-value
	Yes	No		
Age of CSWs (in year)				
15 - 24	55	77	1.00	<.001
25 - 34	36	71	0.71	(X♠ for linear trend)
35 - 49	8	58	0.19	
Educational status of CSWs				
Can't read and write	46	131	1.00	
Can read and write (informal schooling)	14	21	1.90	=.007 (X♠ for linear trend)
Elementary school	22	32	1.96	
High School	17	22	2.20	
Occupation of partners				
Farmer (subsistence)	76	182	1.00	=.023 (X♠ - test)
Trader (Private business)	15	13	2.76	
Others	8	11	1.74	

Helpful tips

a) The **X♠ -test** for linear trend should be considered if a significant finding is obtained by the **ordinary X♠ -test.**

b) Assess the association of variables (dependent and independent) by use of bivariate analysis before going to multivariate analysis.

Exercise

A cross-sectional study to assess the knowledge, attitudes and practice of the population of Dembia district towards traditional harmful health practices was conducted in May 2001. The table below shows the age and sex distribution of the study population and individuals who underwent certain traditional harmful health practices in the last one year preceding the survey [*G.D. Alene and M. Edris, Ethiop.J.Health DEV.2002;16(2):199-07]*.

Variable	study population (n = 6008)	Number of persons who underwent THHP* (n = 368)
Age group (in years)		
< 1	194	41
1 - 4	1255	56
5 - 14	1829	58
15 - 44	2381	184
45 - 64	312	25
65+	37	4
Sex		
Male	2992	195
Female	3016	173

* = traditional harmful health practice

Investigate the impact of age and sex on the experience of the study subjects towards THHPs.

Epi6 performs a lot of tasks using simple statistical techniques. However, it should be noted that Epi6 does not have facilities to perform multivariate analysis, such as, logistic regression. The multiple linear regression technique contained by the analysis program (regress dependent variable = independent variable1, independent variable2, etc.) lacks facilities to check whether the required assumptions are fulfilled. Therefore, it is not possible to control confounding effects with the Epi6 statistical package. In this regard, it will be advisable to consult higher versions of Epi Info and SPSS statistical packages.

Sample size calculation using the Epi Info (STATCALC) program: For population survey or descriptive study

Assumptions:

1) The sample to be taken must be a simple random or otherwise representative sample. A systematic sample, such as every fifth person on a list, is acceptable if the sample is representative.

2) The question being asked must have a "yes/no" or other two-choice answer, leading to a proportion of the population <the "yes's" > as the final result.

Example: Suppose that you wish to investigate whether or not the true prevalence of HIV antibody in a population is 10%. You plan to take a random sample of the population to estimate the prevalence. You would like 95% confidence interval that the true proportion in the entire population will fall within the confidence intervals calculated from your sample.

In STATCALC, therefore, you enter the population size, say 5000, the estimate of the true prevalence (10%), and either 6% or 14% as the "worst acceptable" value, the end point of your sample confidence interval. The program then shows the sample size for several different confidence levels, including the 95% we desired.

The "worst acceptable" value is one of the confidence limits around the estimated sample proportion. The sample size given is for a "two-tailed test", a larger sample size than for a "one-tailed test". The equation works with the following values, if the confidence interval is 95%.

Upper worst acceptable = proportion (P) + (1.96 standard error) = P + marginal error
Lower worst acceptable = proportion (P) - (1.96 standard error) = P - marginal error

In the sample size calculations, an initial screen explains the data items and allows input of a set of values. Pressing <F4> then shows the results (calculated sample sizes) on the second screen.

Sample size calculations for different study designs (more complex designs) are also provided by the STATCALC program.

II. Introduction to SPSS

This section was taken from a resource pack entitled "An Introduction to Practical Statistics Using SPSS" , (Tent Focus, 2002).

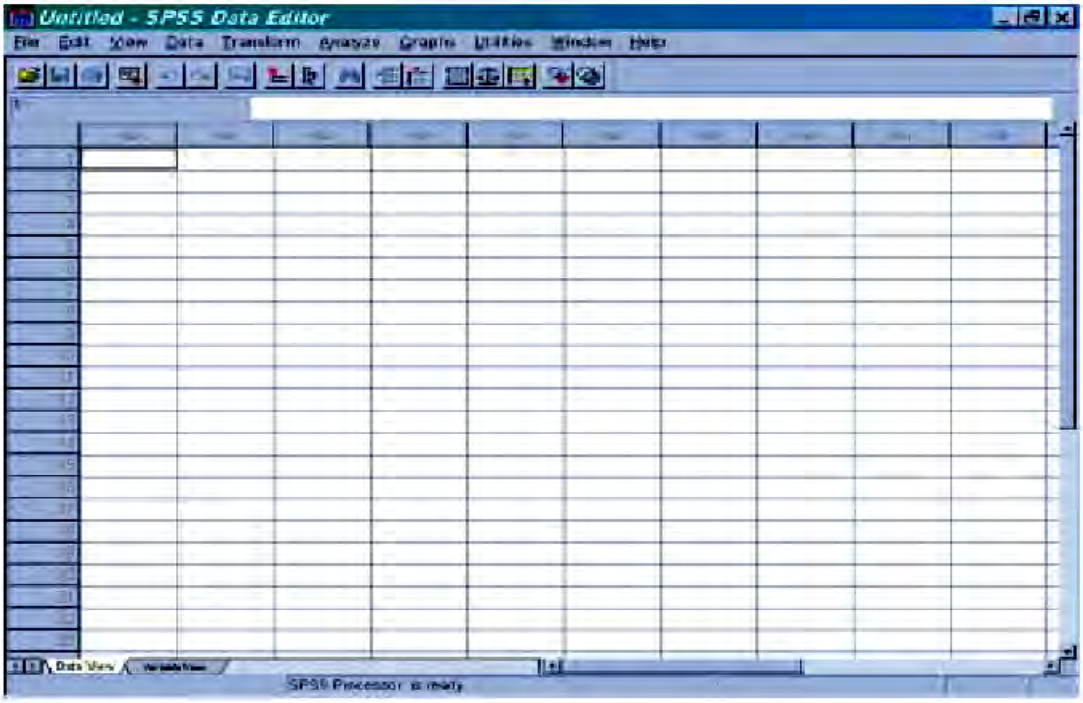
SPSS for Windows: An Introduction

SPSS for Windows is a software package for statistical data analysis. It was originally developed for the Social Sciences, hence the acronym SPSS (**S**tatistical **P**ackage for the **S**ocial **S**ciences), but it is now used in many areas of scientific study. This section serves as an introduction to the package and will show you how to enter and save data. It is written as a step-by-step, interactive exercise, which you can work through if you have access to SPSS for Windows. The data entered will be used in the next Section: Summarising and Presenting Data.

Starting SPSS for Windows

When the program is opened the **SPSS Data Editor** window will appear (Figure 1)

Figure 1 SPSS for Windows[1] Data Editor



The **Menu Bar** displays the names of the menus that are available to you. When you click on a menu name, e.g. File, a list of commands is displayed. The File menu provides options for opening a file, saving a file, printing etc. Other commonly used menus and options are: Edit (to cut, copy or paste data), Analyze (to analyse data e.g. summary statistics, correlation etc.), Graphs (to present data graphically e.g. bar chart, histogram etc.), Window (to move from the Output window to the Data Editor window), and Help. The options on the Menu Bar will depend on which SPSS window is active.

[1] Screen shots from SPSS for Windows Version 10.0.5 are shown throughout this pack (with kind permission of SPSS Inc.).

The **Data Editor window** is similar to a spreadsheet. The rows represent individual cases (observations) and the columns represent variables[2]. A single cell is an intersection of a case and a variable e.g. the height of person x.

The **Output window** is where SPSS displays the statistics and reports from the analysis of your data.

Entering Data

When you start SPSS you are automatically placed in the Data Editor window. The active cell in the window has a heavy black border around it, indicating that any data you type will be entered into that cell. You can move around the Data Editor window by using the arrow keys (←, ↑, →, ↓), or by clicking on cells with the mouse.

Table1 presents some patient data that we can enter in to SPSS. For each patient we have collected the following data from their medical notes: gender, age and blood group. The data has already been coded[3] by a researcher.

Table1: Patient data

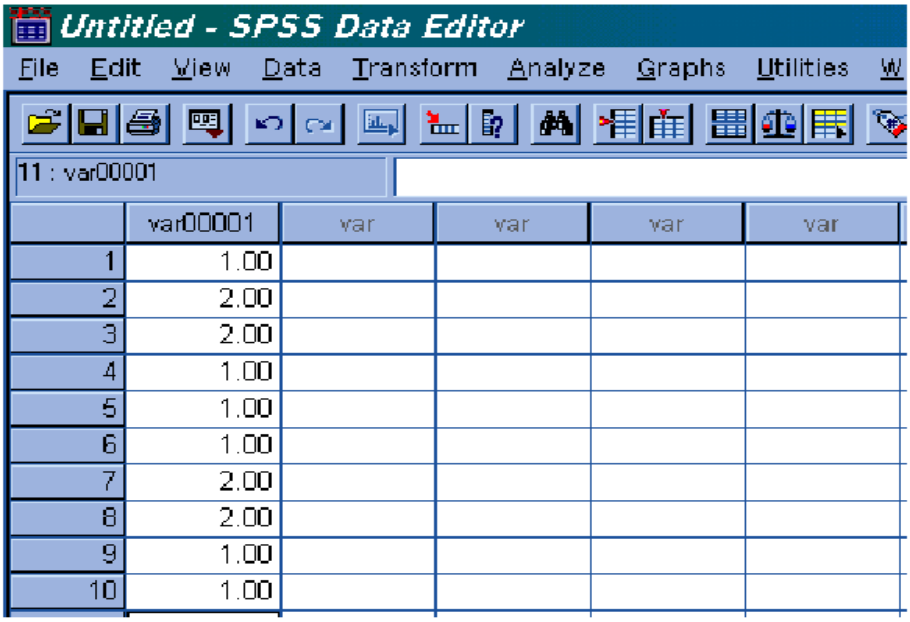
Gender	Age	Blood group
1	21 2	
2	39 1	
2	43 1	
1	55 1	
1	26 4	
1	19 4	
2	65 2	
2	41 2	
1	61 3	
1	50 1	

Where for Gender: 1 = male, 2 = female

and for Blood Group: 1 = O, 2 = A, 3 = AB, 4 = B.

The data can be entered into SPSS as follows:
Starting in the top left cell of the Data Editor window, type in the first value shown above under the column heading Gender, i.e. type 1 and press the **enter** key. This value is then placed in the cell and the black border moves down to the next cell. Type in the other 9 values, remembering to press the **enter** key after each one. Your Data Editor window should look like Figure 2:

Figure 2 Data Editor window



Now move the cursor to the top of the second column and type in the second column of data, Age, and then move to the top of the third column and enter the third column of data, Blood Group. If you make any mistakes during data entry simply return to the incorrect cell using the arrow keys or mouse then type over the entry and press the enter key.

Defining the variables

At the moment the first column is labelled **var0001**. This can be relabelled so that we can give a name to this variable that more easily reminds us what it is. We can also say more about the type of variable we are dealing with (e.g. how many decimal places we want to show, how we want to record missing values) and define the coding scheme we have used. This process is known as **defining the variable** and is described below.

How this is done will depend on which version of SPSS you are using.

(Version 9 or earlier)

First double-click the first grey column heading (currently labelled **var00001**). The **Define Variable** box appears (Figure 3)

Figure 3 Define Variable screen (SPSS Version 9)

(Version 10)

Clicking on the tab labelled **Variable View** at the bottom of the **Data Editor** window brings up a sheet that shows you how each variable is defined and allows you to make changes. Double clicking on any of the column headers on the **Data View** sheet will also bring up this sheet. The **Variable View** sheet is shown in Figure 4.

Figure 4 Variable View sheet

We will amend the information about this variable in several ways:

1. Enter a name for the variable

Variable names (column headings) can be no greater than 8 characters in length and should not contain spaces: hence the name **gender of patient** would not be allowed.

(Version 9 or earlier)

- Type the new name **gender** in the **Variable Name** box
- If we did not wish to make any other changes we would then click on **OK**. However, we shall continue and define the variable further.

(Version 10)

- Type the new name **gender** in the cell in the first row of the **Name** column

2. Enter labels

To overcome the problem of trying to name a variable fully in only 8 characters it is possible to give each variable a label. A label can be up to 256 characters long and include spaces.
To enter labels:

(Version 9 or earlier)

- Click on **Labels** in the Change Settings box
- In the **Variable label** box type **Gender of Patient**

(Version 10)

- Click on the appropriate cell in the **Label** column and type **Gender of patient**

We can also enter the coding scheme that we have used for this variable. This is important for two main reasons. First, it provides us with a permanent record of the scheme. Second it makes interpretation of later analysis much easier as the codes and their meanings are given in full.

(Version 9 or earlier)

- Click in the **Value** box and type **1**
- Click in the **Value Label** box and type **Male**
- Click on the **Add** button
- Click in the **Value** box and type **2**
- Click in the **Value label** box and type **Female**
- Click on the **Add** button
- Click on **Continue**

(Version 10)

- Click on the appropriate cell of the **Values** column and then on the **...** symbol which appears
- In the **Value Labels** window that appears click on the **Value** box and type **1**
- Click in the **Value label** box and type **Male**
- Click on the **Add** button
- Click in the **Value** box and type **2**
- Click in the **Value label** box and type **Female**
- Click on **OK**

3. Change the number of decimal places

By default a column of numeric data will be shown to 2 decimal places and display a maximum of 8 digits for each number. This definition is fine for numbers that require this precision e.g. height of patients in metres such as 1.65, 1.82. However, our data for the gender column requires less precision and can be accommodated in a column with no decimal places i.e. whole numbers. This is done as follows:

(Version 9 or earlier)

- Click on the **Type...** button in the **Change Settings** box
- Click in the **Decimal Places** box and type **0**

(Version 10)

The second column, labelled **type**, shows that the variable **gender** is numeric. The column headed **Width** indicates the number of digits for each number and the column called

Decimals indicates the number of decimal places to be shown

. Click on in the first cell of the fourth column

. Type 0 in this cell, or click on the down arrow twice to reduce the number shown to 0

Before continuing examine the range of data types that are available. Clicking on any cell in the

Type column and then clicking on the **…** symbol will show them.

The Comma, Dot and Scientific Notation options provide for numeric data in different formats. The Date option allows calendar dates to be entered e.g. 24-05-95. The Dollar and Custom currency options provide for currency data e.g. £24.99. The String option is often used to put in comments for each observation; for instance you may wish to record the doctor who saw each patient and enter 'Dr Smith' or 'Dr Jones' as text. The default data format is numeric and most data is best entered in this format, using numeric codes. The use of other codes may restrict the statistical techniques available: for example none of the statistical options at all work with data in string format.

4. Missing values

SPSS is good at handling missing data; there are essentially two options for doing so:

Option 1: System-defined missing values

The system-defined value for missing data in SPSS is '.'. Missing values imported from other software appear on the worksheet as '.', and you can input missing values by entering the full-stop (without quotes) manually.

These missing values are not included in the analysis and SPSS gives a useful summary of missing values and numbers of patients included (the 'case processing summary') as part of the output for all analyses.

However, you may wish to include missing values in some analyses, and it is also good practice to record reasons for missing values if these are known. In this case option 2 is preferable:

Option 2: User-defined missing values

Here codes defined by the user are declared as missing values, using the missing option in the define missing values dialogue (this is accessed via the **define variable** dialogue).

Once all of the variables are correctly defined, clicking on the Data View tab at the bottom of the window will return you to the data sheet.

EXERCISE 1

Define the other two variables, Age and Blood group. For the variable Age use **age** for the column heading; change the type to Numeric8.0; and label the variable as '**Age of patient**'. No value labels are required for the variable as age is used as a continuous measure and is not divided into categories.

For the variable Blood Group use **bloodgp** for the column heading; change the type to Numeric8.0; label the variable as '**Blood group of patient**' and label the values as shown in Table 2.

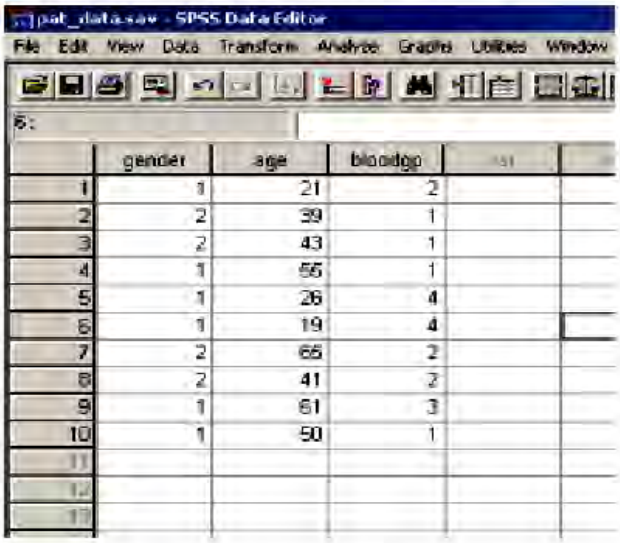
Table 2

Value	Value label
1	O
2	A
3	AB
4	B

Saving your data

The Data View window should look like Figure 5

Figure 5 Data View sheet with names



Now that we have entered the data and defined the properties of all of the variables it is a good time to save the data.

To save the data

. Click on the **File menu**

. Click on **Save As…**

. Once you have chosen the correct drive and directory, type **pat_data** in the **File name** box

. Click on **OK**

SPSS adds the extension **.sav** to the end of your filename, which helps in recognising the file as an **SPSS data file** for future sessions. The data are now ready to be investigated. Examples of analyses using these and other data will be shown later in this pack.

Help

SPSS includes an extensive online help system. The **Help** menu provides different kinds of help, including **Topics**, **Tutorial** and even a **Statistics Coach**. The Statistics Coach may be helpful for choosing the appropriate analysis for a particular dataset. The Help system also includes an online version of the SPSS syntax guide, which is useful for more advanced users.

Exiting SPSS

At any time in your work you may want to exit SPSS. This is achieved with the following menu command:

. Click on the File **menu, then on** Exit**.**

However, do not forget to save your data first if you want to keep the changes you have made.

Concluding Remarks

This Section has introduced you to SPSS and has covered the first steps in using the package to analyse data. This has included entering data, defining variables, saving data and using the Help system.

Both histograms and boxplots can be useful in checking assumptions, such as whether your data are likely to have come from a Normal distribution, and for checking for outliers.

Boxplots are particularly useful, as we shall see later, if you wish to look at more than one group.

Example
This example explains how to summarise and present data using SPSS. The file created in Section 2, **pat_data.sav,** will be used.

If the file is not already open it can be loaded into SPSS as follows:

- Menu commands: **File ⇒ Open...**
- Chose the drive and directory containing the data file
- Click on **pat_data.sav** to put the file name in the **File name** box
- Click on **Open**

Now the dataset is loaded into SPSS we can start looking at it.

(i) Compute summary statistics for **age**

- Menu commands: **Analyze ⇒ Descriptive Statistics ⇒ Descriptives…**
- Click on the variable **age**
- Click on the arrow ↙ to move the variable to the right-hand box
- Click on **OK**
- Inspect the output window:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Age of patient	10	19	65	42.00	16.19
Valid N (listwise)	10				

(ii) Generate a frequency table for **bloodgp**

- Menu commands: **Analyze ⇒ Descriptive Statistics ⇒ Frequencies...**
- Click on the variable **bloodgp** in the left-hand box
- Click on the arrow ↙ to move the variable to the right-hand box
- Click on the button labelled **Statistics...**
- Locate the statistics grouped under the heading **Central Tendency** and click on **Mode** (to find the most common blood group in our sample). The box should be checked with a tick.
- Click on **Continue**
- Click on **OK**
- Inspect the output window:

Statistics

Blood group of patient

N	Valid	10
	Missing	0
Mode		1

Blood group of patient

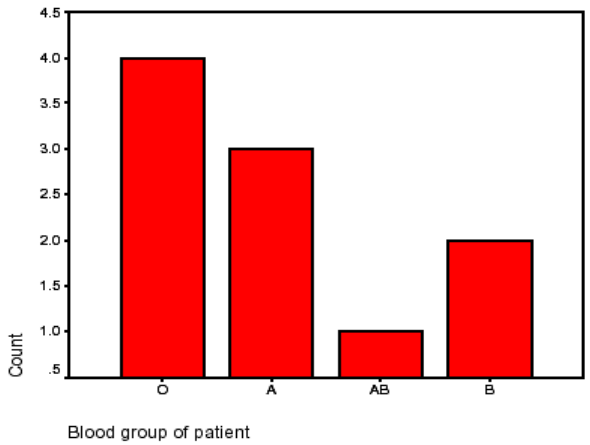
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	O	4	40.0	40.0	40.0
	A	3	30.0	30.0	70.0
	AB	1	10.0	10.0	80.0
	B	2	20.0	20.0	100.0
	Total	10	100.0	100.0	

The first table above shows the number of valid observations and the number of missing observations. This table also shows the mode as being group 1 and we know from Table 2 that this is blood group O. The lower table is a frequency table showing group counts and percentages.

(iii) Generate a bar chart for **bloodgp**

- Menu commands: **Graphs ⇒ Bar...**
- Click on **Simple** and then on the button labelled **Define**
- Click on the variable **bloodgp**
- Click on the arrow ↙ next to the label **Category Axis:** to move the variable into the box. Ensure that **Bars Represent N of cases is checked**
- Click on **OK**

Figure 9 *Bar chart for blood group*



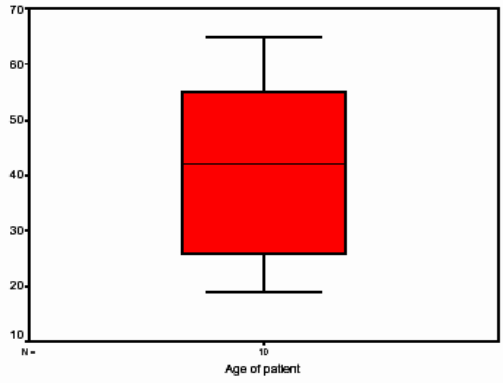
(iii) Generate a boxplot for **age**

- Menu commands: **Graphs ⇒ Boxplot...**
- Ensure that the box labelled **Simple** is highlighted
- Ensure that **Summaries of separate variables** is checked
- Click on **Define**
- Click on the variable *age*
- Click on the arrow ↙ next to the label **Boxes represent:** to move the variable into the box
- Click on **OK**

Explore

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Age of patient	10	100.0%	0	.0%	10	100.0%



EXERCISE 2

Generate a histogram for **age.**

Concluding Remarks

In this Section we have look at a few ways of summarising and displaying data. This is an important part of any analysis but is often overlooked in the rush to apply more complex methods. However time spent understanding your data and checking for potential problems is likely to save you time and difficulties later.

Double clicking on graphics in the **Output Window** will open the **Chart Editor Window** and allow you to customise your graphics. It is possible, for example, to change symbols, colours and line types as well as to define and label axes.

9.3 Exporting files from Epi6 to Epi Info 2000 and SPSS

I) Export files from Epi6 to Epi info 2000

- There are two ways of exporting **Rec** files from Epi6 to Epi info 2000 (we will show here only one of them). This is usually done to analyze data using advanced statistical techniques, such as, logistic regression.
- Move the cursor to **Epi info 2000** and click on it.
- Move the cursor to **Adv. Stats** and click on **logistic regression.**
- Move the cursor to **File** and click on **open**
- Make sure that the current drive is **"C"** (if this is the drive where your Epi6 rec file is found. Otherwise, you need to specify the correct drive.
- **List files of type:** move the cursor to Data file (*.rec) and press Enter
- Find the required file from the list of files given above and click on it.
- If you don't find the required "rec" file from list of files given above, you should write the appropriate drive, directories and the name of the file on the space provided below "**File name"** and press Enter.
- Example: **C:\Epi6\Trach33.rec**
- Now, move the cursor to "**Model**" and click on "**Construct model**"
- Move the cursor to "**Outcome**" and click on the down arrow below it.
- Choose the outcome (dependent) variable and click on it.
- Move the cursor to **variables :**

  - put the cursor on the variable (independent) which is assumed to be a predictor of the outcome variable and click on the " **+** " sign.

- Do the same thing until all the required independent variables are considered.

- Click on **OK**
- Move the cursor to **Statistics** and click on **unconditional**
- **Did you get the logistic regression model ?**

II) Export files from Epi6 to SPSS

Data on Epi6 cannot be exported to SPSS directly. First, it should be exported to Dbase 4 and the newly created Dbase file will be exported to SPSS. The procedure is given as follows:

- Open the Epi6 program
- Move the cursor to Export files and press Enter
- **Input file Name**: Give the input file name (The epi6 file which you would like to export to SPSS) - Don't forget to include the extension **Rec**.

  Example: **Bekele22.rec**

- Move the cursor to **dbase 4** and press the Enter key
- Output file name: Give a file name ending in **DBF** and press the Enter key.

  Example: **Bekele22.DBF**

- Now, make sure that the following statement appears
- **Exported ____ records from Bekele22.rec to bekele22.dbf**

(The blank space indicates the number of records entered using the Epi6 program**)**

- **Close the Epi6 program**
- Click on the **SPSS** 10 icon (or any higher version of SPSS)

- Click on cancel
- Click on File
- Move the cursor to **open** and click on **Data**
- **Look in: SPSS** (press the down arrow and click on **Local disk (c:)**
- move the cursor to **Epi6** and click on **open**
- **Files of type:** (press the down arrow and click on **dbase (*.dbf)**
- Now, the required file appears below the **EPI** directory and click on it.
- click on **open**
- At this stage, you should see some descriptions regarding the type of variables and the number of cases (records)
- Move the cursor to **Untitled SPSS Data** and click on it.
- Delete any unnecessary records at the first row (A1) (if any)
- Finally, save the file as an SPSS data file with extension **.sav**

REFERENCES

1. Corlien M. Varkevisser, Indra Pathmanathan, and Ann Brownlee. Designing and conducting health systems research projects: Volume 1 Proposal development and fieldwork. KIT/IDRC. 2003
2. Degu G, Tessema F. Biostatistics for Health Science Students: lecture note series. The Carter Center 9EPHTI), Addis Ababa; January 2005.
3. Abramson JH. Survey methods in community medicine. 2nd ed. Eidenburgh: Churchill Livingstone, 1979.
4. Altman DG. Practical Statistics for Medical Research. London: Chapman and Hall, 1991.
5. Colton T. Statistics in Medicine. Boston: Little, Brown and Company 9INC.), November 19974.
6. Mathers, Nigel; Howe, Amanda; and Hunn Amanda. Trent focus for research and development in primary health care. Ethical considerations in research. Trent focus, 1998
7. ESTC-EPHA/CDC PROJECT. Training modules on health research. 2004.
8. Department of Community Health ,Jimma Institute of Health Sciences. Manual for student research project. Jimma, April 1996.
9. Department of Community Health, Gondar College of Medical Sciences. Manual for field training. Gondar, 1995.
10. Department of Community Health, Faculty of Medicine. Handout for Rural Community Health Training Programme. January 2002.
11. Fletcher M. Principles and Practice of Epidemiology. Department of Community Health, Faculty of medicine, Addis Ababa University. August 1992.
12. Manktelow B, Hewitt MJ, Spiers N. An introduction to Practical Statistics Using SPSS. Trent Focus, 2002.