LECTURE NOTES

Biostatistics

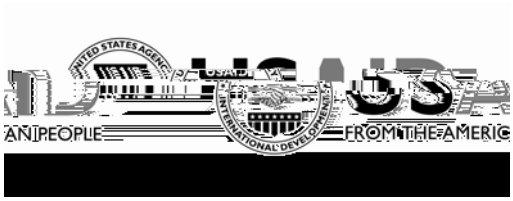
For Health Extension Workers

Getu Degu

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

November 2004



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.

Acknowledgments

The development of this lecture note for training Health Extension workers is an arduous assignment for Ato Getu Degu at Gondar University.

Essentially, it required the consolidation and merging of existing in depth training materials, examination of Health Extension Package manuals and the Curriculum.

Recognizing the importance of and the need for the preparation of the lecture note for the Training of Health Extension workers THE CARTER CENTER (TCC) ETHIOPIA PUBLIC HEALTH TRAINING INITIATIVE (EPHTI) facilitated the task for Gondar University to write the lecture note in consultation with the Health Extension Coordinating Office of the Federal Ministry of Health.

Finally the Federal Ministry of Health would like to express special words of gratitude for those who contributed and endeavored to the development of this lecture note and to TCC/USAID for the technical and financial support.

Table of Contents

Acknowledgements i
Table of contents ii
Introduction 1

UNIT ONE: Introduction to Statistics 3

1.1 Learning objectives 3
1.2 Introduction 3
1.3 Definition of terms 10

UNIT TWO: Collection of Statistical Data 13

2.1 Learning objectives 13
2.2 Types of data 13
2.3 Data collection 18

UNIT THREE: Methods of Data Presentation 27

3.1 Learning objectives 27
3.2 Introduction 27
3.3 Tabular presentation of data 28
3.4 Diagrammatic representation of data 38

UNIT FOUR: Demographic Methods 53

4.1 Learning objectives 53
4.2 Introduction 53
4.3 Sources of demographic data 54

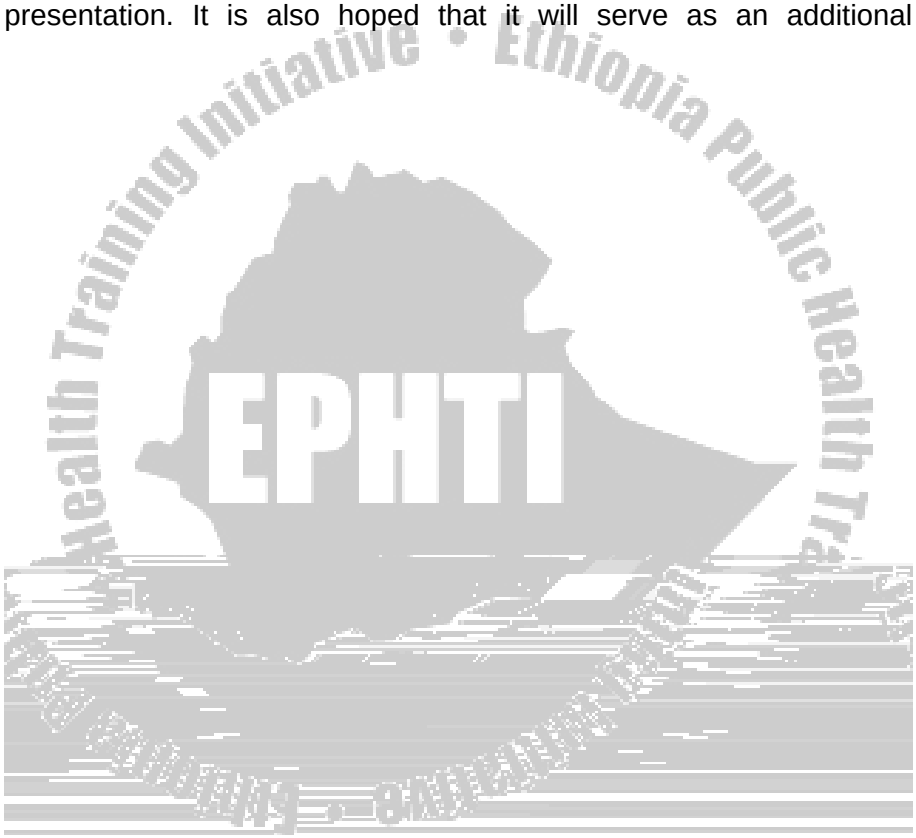
UNIT FIVE: Measures of Central Tendency and Variation 61

5.1 Learning objectives 61
5.2 Introduction 61
5.3 Measures of central tendency 63
5.4 Measures of variation 73

References 82

Introduction

This lecture note is prepared primarily for the health extension workers who need to know the basic principles of data collection and presentation. It is also hoped that it will serve as an additional



questions, most of them based on real data. A few reference materials are given at the end of the lecture note for further reading.

UNIT ONE

Introduction to Statistics

1.1. Learning Objectives

After completing this unit, the trainee will be able to:

- Define statistics
- Enumerate the importance of statistics
- Understand the limitations of statistics

1.2. Introduction

This section deals with the basic concepts and definitions of statistics and related terms.

Definition: The term statistics is used to mean either statistical data or statistical methods.

A. Statistical data: When it means statistical data it refers to numerical descriptions of things. These descriptions may take the form of counts or measurements. Thus statistics of malaria cases in one of the health posts of Ethiopia include fever cases, number of positives obtained, sex and age distribution of positive cases, etc.

NB Even though statistical data always denote figures (numerical descriptions) it must be remembered that all 'numerical descriptions' are not statistical data.

Characteristics of statistical data

In order that numerical descriptions may be called statistics they must possess the following characteristics:

- They must be in aggregates – This means that statistics are 'number of facts.' A single fact, even though numerically stated, cannot be called statistics.
- They must be affected to a marked extent by a multiplicity of causes – This means that statistics are aggregates of such facts only as grow out of a ' variety of circumstances'. Thus the explosion of an outbreak of malaria is attributable to a number of factors, Viz., Human factors, parasite factors, mosquito and environmental factors. All these factors acting jointly determine the severity of the outbreak and it is very difficult for any one to assess the individual contribution of any one of these factors.
- They must be enumerated or estimated according to reasonable standard of accuracy. This means that if aggregates of numerical facts are to be called 'statistics' they must be reasonably accurate. This is necessary because statistical data are to serve as a basis for statistical investigations. If the basis happens to be incorrect the results are bound to be misleading.
- They must have been collected in a systematic manner for a predetermined purpose. Numerical data can be called statistics only if they have been compiled in a properly

Health Training Initiative • Ethiopia Public Health Training
EPHTI

1. It presents facts in a definite form. Statements (facts) given numerically are definite and hence more convincing than facts stated qualitatively.

Example: a) We have recorded more malaria patients this year than the previous year.

b) We have recorded 2500 malaria patients this year compared to 1500 of the previous year.

The facts given in part 'b' are definite and more convincing.

2. Statistics simplifies huge and complex mass of data. The raw data are usually huge and should be reduced to some simpler form so that we can understand the main features of the data very easily. The complex data may be reduced to totals, averages, percentages, etc. and presented either in tabular or diagrammatic forms. For example, if the list of the ages of patients who visited a given health facility in the last ten years is given, it will be difficult to understand the age groups which were highly affected. Therefore, these data should be reduced to something simpler (eg., average age) so that we can easily point out the most affected age group.

3. Statistics classifies numerical facts. The procedure of classification helps us to have a better understanding of the variable

Example: a) the variable sex could be classified as male or female.

b) marital status could be classified as single, married, divorced or widowed.

c) etc.

4. Statistics furnishes a technique of comparison. The facts, having been once classified, are now in a shape when they can be used for purposes of comparisons. Certain facts, by themselves, may be less meaningful unless they are capable of being compared with similar facts at other places or at other periods of time.

5. Statistics endeavours to interpret conditions. Based on the existing facts (eg. disease conditions, access to safe water, availability of latrines, etc.) we can interpret the situation to a greater extent and develop mechanisms which would alleviate the given problem.

D. Importance of statistics: The need for statistics in the smooth functioning of an undertaking in any sector is very high. Students who are interested to know the uses of statistics in many other fields could refer to books listed at the end of this module. The uses of statistics in the health service are described below.

Health Service Statistics: Health statistics are very useful to improve the health situation of the population of a given country. For example, the following questions could not be answered correctly unless the health statistics of a given area is consolidated and given due emphasis.

a) What is the leading cause of death in the area? Is it malaria, tuberculosis, etc.?

b) At what age is the mortality highest, and from what disease?

c) Are certain diseases affecting specified groups of the population more than others? (This might apply, for example, to women or children, or to individuals following a particular occupation.)

d) In comparison with similar areas, is this area healthier or not?

e) Are the health institutions in the area able to cope with the disease problem?

f) Is there any season at which various diseases have a tendency to break out? If so, can these be distinguished?

g) What are the factors involved in the incidence of certain diseases, like malaria, tuberculosis, etc.?

The functions/uses of health statistics are enormous. A short list is given below.

Health service statistics are used to:

- describe the level of community health
- diagnose community ills
- discover solutions to health problems and find clues for administrative action
- determine priorities for health programmes
- promote health legislation
- determine the met and unmet health needs
- disseminate information on the health situation and health programmes
- determine success or failure of specific health programmes
- demand public support for health work

E. limitations of statistics: Some of the important limitations are given below.

- It deals with only those subjects of inquiry that are capable of being quantitatively measured and numerically expressed.

- Á It deals with aggregates of facts and no importance is attached to individual items .
- Á Statistical data are only approximately and not mathematically correct. For example, the age of an individual could be 40 years and 8 months and 10 days, etc. However, we use an approximate value of 41 years.

1.3. Definition of terms

- **Biostatistics** - When the different statistical methods are applied in biological, medical and public health data they constitute the discipline of biostatistics**.**
- **Descriptive statistics -** One branch of statistics which deals with the description of data in a clear and informative manner using tables and graphs. Also, it refers to the methods used to summarize a body of data with one or two meaningful figures. These are the types of statistics most commonly heard over the radio.
- **Vital statistics** - One branch of descriptive statistics of special relevance in public health is that of vital statistics (the recording of vital events as they occur). The most important vital events are: births, deaths, marriages, divorces, migration and the occurrence of particular diseases. They are used to characterize the health status of a population. Coupled with

results of periodic censuses and other special enumeration of populations, the data on vital events relate to an underlying population and yield descriptive measures such as birth rates, morbidity rates, mortality rates, life expectancies, and disease incidence and prevalence rates.

- **population (also called source population or target population or reference population) -** the entire group of interest, to which the investigators would like to generalize the results of the study, and from which a representative sample is to be drawn.
- **sample** - A sample is a part of the population.

Example: A representative sample of 400 under-five children was taken from a population of 2500 under-five children in a certain district to study the level of malnutrition of the area.

Exercises

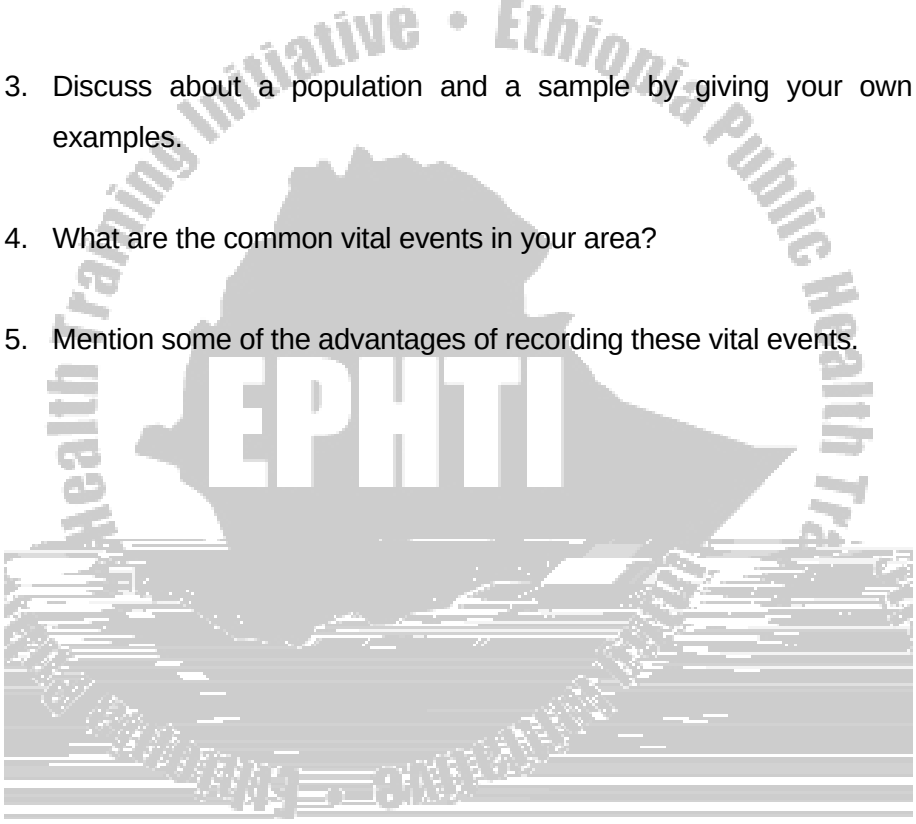
1. What is the difference between statistics and biostatistics?

2. Write a short note on the uses of health service statistics.

3. Discuss about a population and a sample by giving your own examples.

4. What are the common vital events in your area?

5. Mention some of the advantages of recording these vital events.



UNIT TWO

Collection of Statistical Data

2.1. Learning Objectives

At the end of this unit, the trainee will be able to:

- Á enumerate the various types of data (scales of measurement).
- Á describe the most important sources of data.
- Á describe the main methods of data collection.

2.2. Types of data (Scales of Measurement)

Any thing (phenomenon) which takes different values for different individuals or cases, like marital status, blood pressure, age, sex, etc. vadri abl.(describe the)50779 0ls

discrete or continuous. The values of a discrete variable are usually whole numbers, such as the number of episodes of diarrhoea in the first five years of life. A continuous variable is a measurement on a continuous scale. Examples include weight, height, blood pressure, age, etc.

Although the types of variables could be broadly divided into categorical (qualitative) and quantitative, it has been a common practice to see four basic types of data (scales of measurement).

Nominal data:- Data that represent categories or names. There is no implied order to the categories of nominal data. In these types of data, individuals are simply placed in the proper category or group, and the number in each category is counted. Each item must fit into exactly one category. Marital status is an example of nominal data. The categories are: single, married, divorced and widowed. There is no order in the arrangement of the categories of marital status. That is, it is also possible to write the categories of marital status in the following arrangement; married, single, divorced and widowed.

Among the various nominal data, the simplest types consist of only two possible categories. These types of data are called dichotomous. That is, either the patient lives or the patient dies, either he/she has some particular attributes or he/she does not. An attribute is a characteristic

that an individual possesses. The categories (characteristics) given above could be taken as attributes.

The example below shows the nominal scale data which is dichotomous.

Survival status of propranolol - treated and control patients with myocardial infarction (**MI).**

Status 28 days after hospital admission	Propranolol -treated patients	Control Patients
Dead	7	17
Alive	38	29
Total	45	46
Survival rate	84%	63%

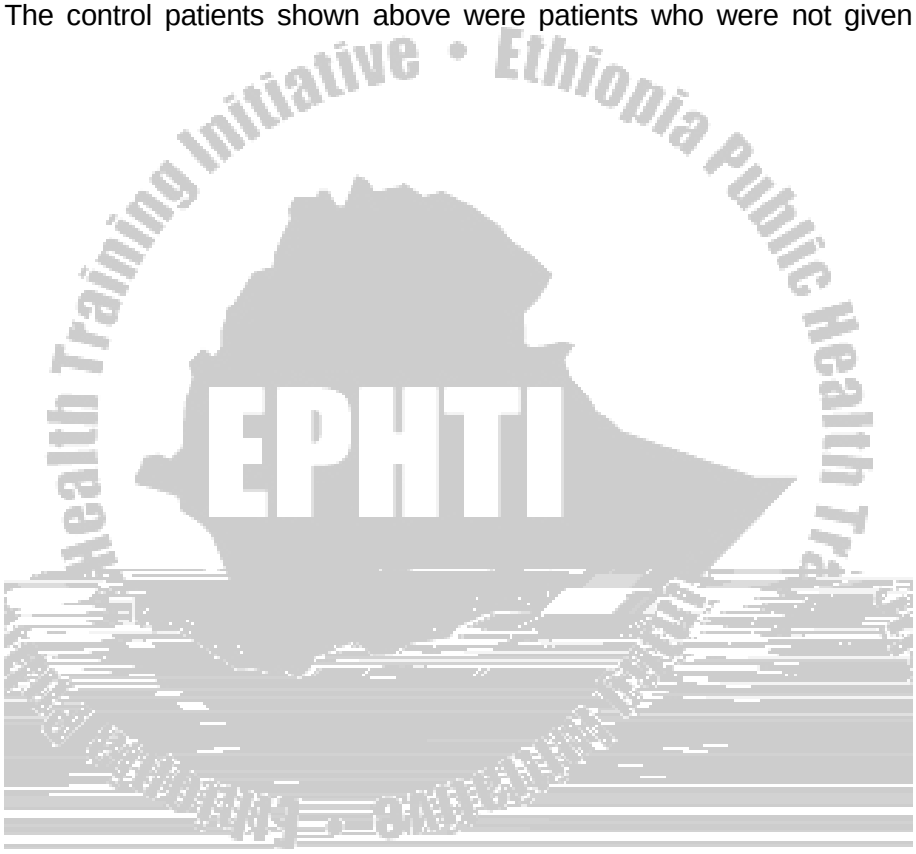
Source: Snow, effect of propranolol in MI ;The Lancet, 1965.

The above table presents data from a clinical trial of the drug propranolol in the treatment of myocardial infarction. There were two group of patients with MI. One group received propranolol; the other did not and was the control. For each patient the response was dichotomous; either he/she survived the first 28 days after hospital

admission or he/she succumbed (died) sometime within this time period.

Propranolol is a drug used to treat myocardial infarction (MI).

The control patients shown above were patients who were not given



2.3. Data Collection

a) Sources of data

The statistical data may be classified under two categories, depending upon the sources.

1) Primary data and Secondary data

Primary Data: are those data, which are collected by the investigator himself/herself for the purpose of a specific inquiry or study. Such data are original in character and are mostly generated by surveys conducted by individuals or research institutions.

The first hand information obtained by the investigator is more reliable and accurate since the investigator can extract the correct information by removing doubts, if any, in the minds of the respondents regarding certain questions. High response rates might be obtained since the answers to various questions are obtained on the spot. It permits explanation of questions concerning difficult subject matter.

Secondary Data: When an investigator uses data, which have already been collected by others, such data are called "Secondary Data". Such data are primary data for the agency that collected them, and become secondary for someone else who uses these data for his/her own purposes.

The secondary data can be obtained from journals, reports of different institutions, government publications, publications of professionals and research organizations. These data are less expensive and can be collected in a short time.

On the other hand, such data must be used with great care, because such data may also be ***full of errors*** due to the fact that the purpose of the collection of the data by the primary agency may have been different from the purpose of the user of these secondary data. Moreover, there may have been bias introduced, the size of the sample may have been inadequate, or there may have been arithmetic or definition errors, hence, it is necessary to critically investigate the validity of the secondary data.

The selection of a particular source is dependent upon a variety of factors, such as:

- purpose of the inquiry
- time period
- accuracy desired
- funds available
- other facilities available (transport, etc,)

The most important thing that should be noted is that the work of collecting facts should be undertaken in **a planned manner**.

b) Methods of data collection

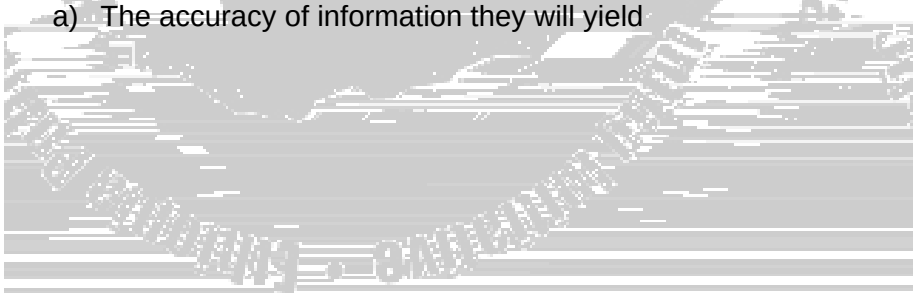
Depending on the type of variable and the objective of the study different data collection methods can be employed.

Data collection techniques allow us to systematically collect data about our objects of study (people, objects, and phenomena) and about the setting in which they occur. In the collection of data we have to be systematic. If data are collected haphazardly, it will be difficult to answer our questions in a conclusive way.

The methods of collecting information may be broadly classified as: observation, the documentary sources, interviews and self-administered questionnaires.

The choice of methods of data collection is based on:

a) The accuracy of information they will yield



measurements, sophisticated equipment or facilities, such as radiographic, biochemical, X-ray machines, microscope, clinical examinations, and microbiological examinations.

Advantages: Gives relatively more accurate data on behaviour and activities

Disadvantages: Investigators or observer's own biases, prejudice, desires, and etc. and needs more resources and skilled human power during the use of high level machines.

2. **The Documentary sources:** Documentary sources include clinical records and other personal records, published mortality statistics, census publications, etc.

Advantages: a) Documents can provide ready-made information relatively easily

b) The best means of studying past events

Disadvantages: a) 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).

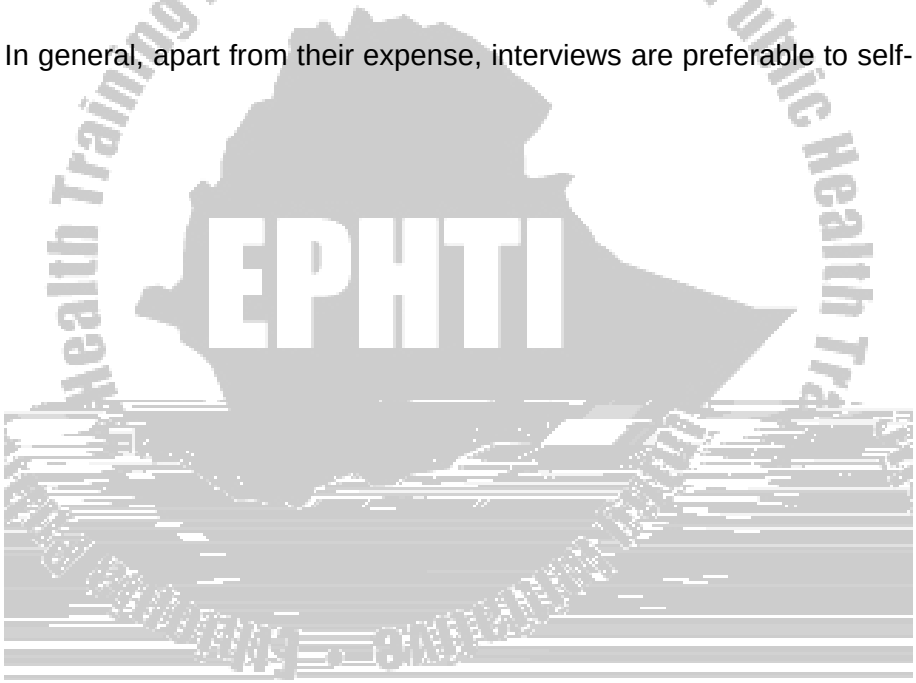
b) There is a possibility that errors may occur when the information is extracted from the records (this may be an important source of unreliability if handwritings are difficult to read).

the questionnaire, ask them questions and record their replies. This can be done using telephone or face-to-face interviews.

Questions may take two general forms: they may be “open ended”

- 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/she says it.

In general, apart from their expense, interviews are preferable to self-



Checklist - is a list of questions prepared ahead of time to facilitate the interviews or discussions. It is not an exhaustive one. It helps the facilitator not to miss any of the important topics under consideration.

Key informant interviews – interviews done with influential individuals (such as community elders, priests, etc.).

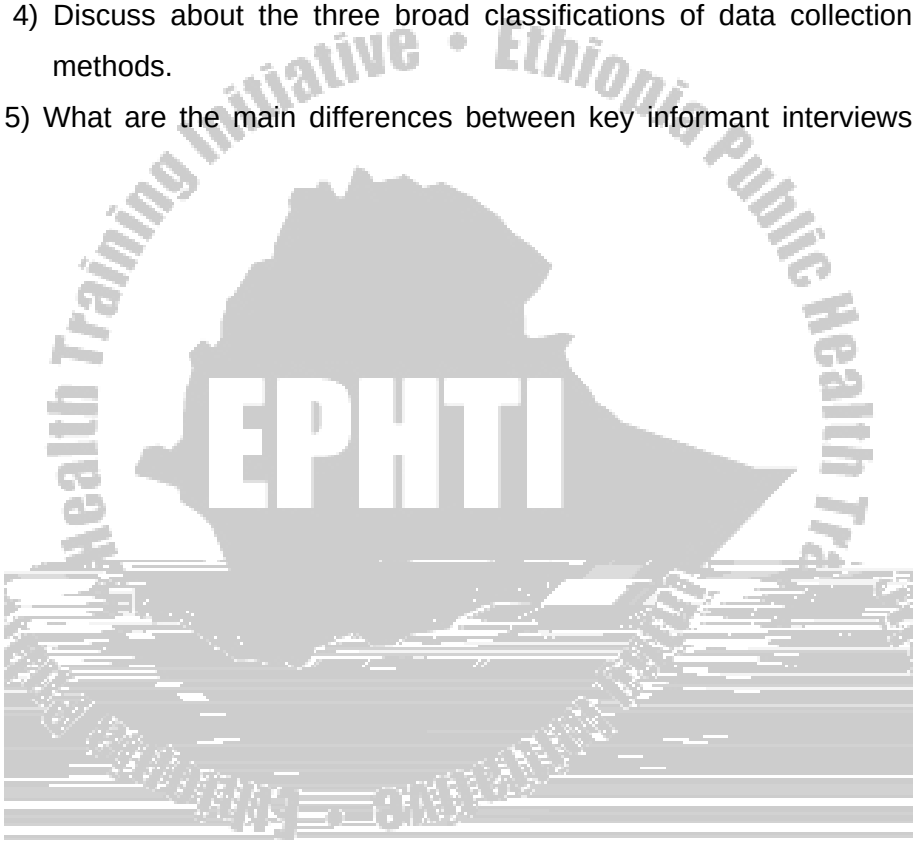
Focus group discussions – discussions made with a group of respondents. The group contains 6 to 12 people who are more or less similar with respect to level of education, marital status, age, sex, etc. (this composition helps each respondent to talk freely without being dominated by the other).

Exercises

1) Identify the type of data (nominal, ordinal, interval and ratio) represented by each of the following. Confirm your answers by giving your own examples.

a) Blood group
b) Temperature (Celsius)
c) Ethnic group
d) Job satisfaction index (1-5)
e) Number of heart attacks
f) Serum uric acid (mg/100ml)
g) Number of accidents in 3 - year period
h) Number of cases of each reportable disease reported by a health worker

2) What are the strengths and limitations of primary data and secondary data?
3) The work of data collection should be done on a planned manner. Why ?
4) Discuss about the three broad classifications of data collection methods.
5) What are the main differences between key informant interviews



UNIT THREE

Methods Of Data Presentation

3.1. Learning Objectives

At the end of this unit, the trainee will be able to:

- describe the types of tabular presentations
- construct statistical tables
- enumerate the most important types of graphs
- understand the basic features of the different graphs

3.2. Introduction

The data collected in a survey is called *raw data*. In most cases, useful information is not immediately evident from the mass of these raw data. Collected data need to be organized in such a way as to condense the information they contain in a way that will show patterns of variation clearly. Therefore, for the raw data to be more easily appreciated and to draw quick comparisons, it is often useful to present the data in the form of:

- Ordered arrays – If the number of observations is not too large (usually less than 20), a first step in organizing these data is the preparation of an ordered array. It is an arrangement of the figures in increasing or decreasing order.
- Tables – An orderly and systematic presentation of data in rows and columns.

- Graphs – diagrammatic representation

3.3. Tabular presentation

As indicated above, tabular presentation refers to the systematic arrangement of data in rows and columns. A table should not be a misleading one. It should present a truthful impression of the data.

frequency table (distribution) - A distribution of observations in the form of tables showing frequencies. A frequency table contains only one variable.

cross-tabulation – A frequency table involving at least two variables that have been cross-tabulated.

Parts of a table

a) **Title** : it explains

- what the data are about
- from where the data are collected
- time period when the data are collected
- how the data are classified

b) **Captions**

The headings of the **columns** are given in captions. In case there is a sub-division of any column, there would be sub-caption headings too.

c) **Stubs** : the headings of **rows** are called stubs.

d) **Body** : contains the numerical data.

e) **Head note** : a statement below the title which clarifies the content of the title.

f) **Foot note**: a statement below the table which clarifies some specific items given in the table. For example, it explains omissions, etc. in the table.

g) **Source**

The average number of cups of caffeinated coffee taken per day by each person	Frequency (number of persons)	Relative frequency (%)
0	20	5.0
1	40	10.0
2	100	15.0
3	150	37.5
4	90	22.5
5	35	7.5
6	15	2.5
Total	450	100.0

In the above distribution the average number of cups of caffeinated coffee represents the variable under consideration, number of persons represents the frequency, and the whole distribution is called a frequency distribution.

Grouped frequency distribution

Consider the problem of an investigator who wants to study the ages of persons who had car accidents during one year in a country. In connection with large sets of data, a good overall picture and sufficient information can often be conveyed by grouping the data into a number of class intervals as shown below.

Age (years)	Number of persons
18 – 24	4860
25 – 34	3240
35 – 44	1620
45 – 54	756
55 and over	324
Total	10800

This kind of frequency distribution is called grouped frequency distribution.

Frequency distributions present data in a relatively compact form, give a good overall picture, and contain information that is adequate for many purposes, but there are usually some things which can be determined only from the original data. For instance, the above grouped frequency distribution cannot tell how many of the persons were 20 years old, or how many were over 60.

The construction of grouped frequency distribution consists essentially of four steps:

(1) Choosing the classes, (2) sorting (or tallying) of the data into these classes, (3) counting the number of items in each class, and (4) displaying the results in the form of a chart or table

Choosing suitable classification involves choosing the number of classes and the range of values each class should cover, namely, from where to where each class should go. Both of these choices are arbitrary to some extent, but they depend on the nature of the data and its accuracy and on the purpose the distribution is to serve. The following are some rules that are generally observed:

1) Determine the number of classes
   - should lie between 6 and 20
   - should accommodate your data
   - should be mutually exclusive
   - whenever possible make the class intervals equal

A guide on the determination of the number of classes (k) can be the Sturge's Formula, given by:

K = 1 + 3.322×log(n), where n is the number of observations.

Example: a) log 100 =2

b) log 150 = 2.18

c) log 80 = 1.9

The length or width of the class interval (w) can be calculated by:

W = (Maximum value – Minimum value) / K = Range / K

Range = maximum value – minimum value

2) Determination of class limits
   -

Health Training Initiative • Ethiopia Public Health Tra
EPHTI

29	30	35	37	40	47	50	65	78	87
71	75	64	51	47	41	38	35	37	26
63	60	52	53	48	49	42	43	31	39
44	33	36	38	27	25	36	32	36	35
54	55	56	69	57	47	47	45	46	31
47	48	59	58	21	23	22	32	31	30
48	49	32	33	34	39	33	34	23	22
40	41	49	48	42	43	47	46	43	44

Using the above formula, K = 1 + 3.322 $\times$ log (80) = 7.32 $\approx$ 7 classes

Maximum value = 87 and Minimum value = 20 Č Range = 87 – 20 = 67

and W (by following the formula given above) = 67/7 = 9.6 $\approx$ 10

Using a width of 10 and by applying the procedure indicated above,

Note that the unit of measure is 1. You can find this value by taking the difference of two consecutive numbers (i.e., consider any two numbers with the smallest difference, like 42-41, 22-21, etc.).

In the case of large size quantitative variables like weight, height, etc. measurements, the groups are formed by amalgamating continuous values into classes of intervals. There are, however, variables which have frequently used standard classes. One of such variables, which have wider applications in demographic surveys, is age. The age distribution of a population is described based on the following intervals:

< 1	20-24	45-49
1-4	25-29	50-54
5-9	30-34	55-59
10-14	35-39	60-64
15-19	40-44	65+

Based on the purpose for which the table is designed and the

Number of persons with malaria parasites obtained from outbreak reports of 1973-1978 (Ethiopia) classified by species type.

Year	Total positives	Species distribution			
		P.f.	P.v.	P.m.	mixed (P.f.+P.v.)
1973	21969	15214	6519	23	213
1974	5343	2911	2419	13	0
1975	498	361	126	7	4
1976	4318	2566	1638	5	109
1977	4988	3310	1570	9	99
1978	10859	7299	3361	8	191
Total	47975	31661	15633	65	616

P.f. = Plasmodium falciparum, P.v. = Plasmodium vivax, P.m . = Plasmodium malariae

Source: Malaria and other vector borne diseases control organization, Ethiopia, 1978.

Importance of statistical tabulation

- tabulated data can be easily understood than facts stated in the form of description
- they facilitate comparison
- they leave a lasting impression
- they make easier the summation of items
- detection of errors and omissions is facilitated
- when data are tabulated all unnecessary details and repetitions are avoided.

Construction of tables

Although there are no hard and fast rules to follow, the following general principles should be addressed in constructing tables.

1. Tables should be as simple as possible.
2. Tables should be self-explanatory. For that purpose
    - title should be clear and to the point(a good title answers: what? when? where? how classified ?) and it be placed above the table.
    - each row and column should be labelled.
    - numerical entities of zero should be explicitly written rather than indicated by a dash. Dashed are reserved for missing or unobserved data.

- totals should be shown either in the top row and the first column or in the last row and last column.

3. If data are not original, their source should be given in a foot note.
4. Overall, tables should be clearly labelled and the reader should be able to determine without difficulty precisely what is tabulated.

3.4. Diagrammatic Representation of Data

Appropriately drawn graph allows readers to obtain rapidly an overall grasp of the data presented. The relationship between numbers of various magnitudes can usually be seen more quickly and easily from a graph than from a table.

Figures are not always interesting, and as their size and number increase they become confusing and uninteresting to such an extent that no one (unless he/she is specifically interested) would care to study them.

Importance of Diagrammatic Representation

- They have greater attraction than mere figures.
- They help in deriving the required information in less time and without any mental strain.
- They facilitate comparison.

Á

Health Training Initiative • Ethiopia Public Health Training
EPHTI

- A proper scale should be selected and the units in to which the scale is divided should be clearly indicated.
- The vertical and horizontal scales should be clearly shown on the diagram itself - the former on the left hand side and the latter at the bottom of the diagram.
- The numerical scale representing frequency must start at zero or a break in the line should be shown.
- Titles of diagrams should be self explanatory. That is, the type of data (what), place that the data were collected (where), time period (when) and How the data were classified should be shown.
- Diagrams should be as simple as possible.
- Legends or keys should be used to differentiate variables if more than one is shown.
- Neatness should be strictly observed.
- Source should be given (if data are not original).

N.B. The graph (diagram) should not be a misleading one. It should present a truthful impression of the data.

Among the kinds of diagrams in common use are:

1. Bar graph

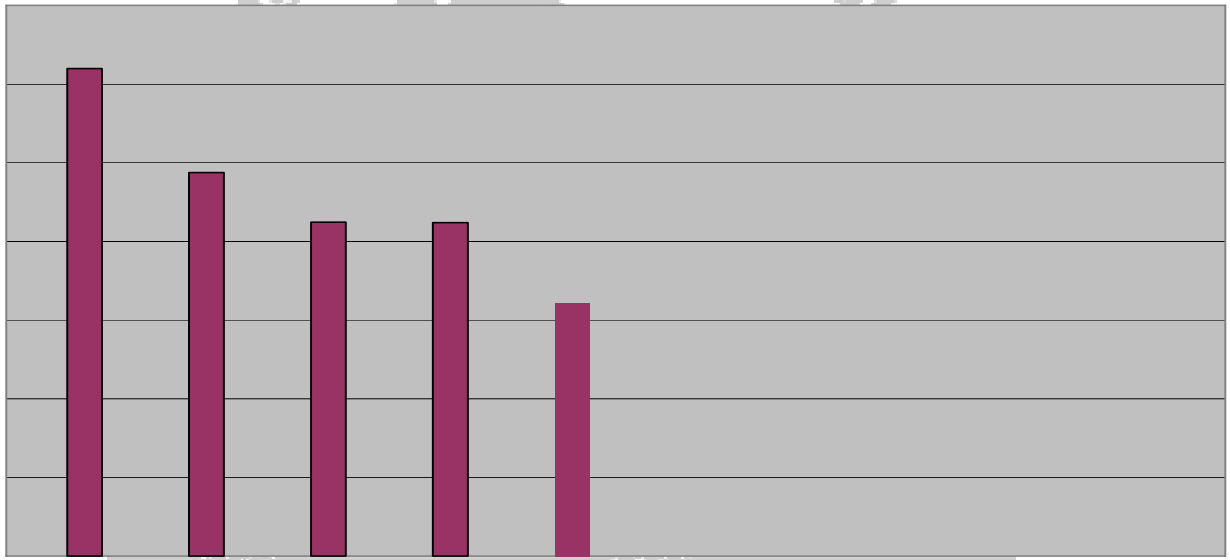
Bar diagrams are used to represent and compare the frequency distribution of discrete variables and attributes of categorical series.

When we represent data using bar diagram, all the bars must have **equal width** and the **distance** between bars must be **equa**l.

There are different types of bar diagrams, the most important ones are:

a) ***Simple bar chart:*** It is a one-dimensional diagram in which the bar represents the whole of the magnitude. The height or length of each bar indicates the size (frequency) of the figure represented.

Example 1

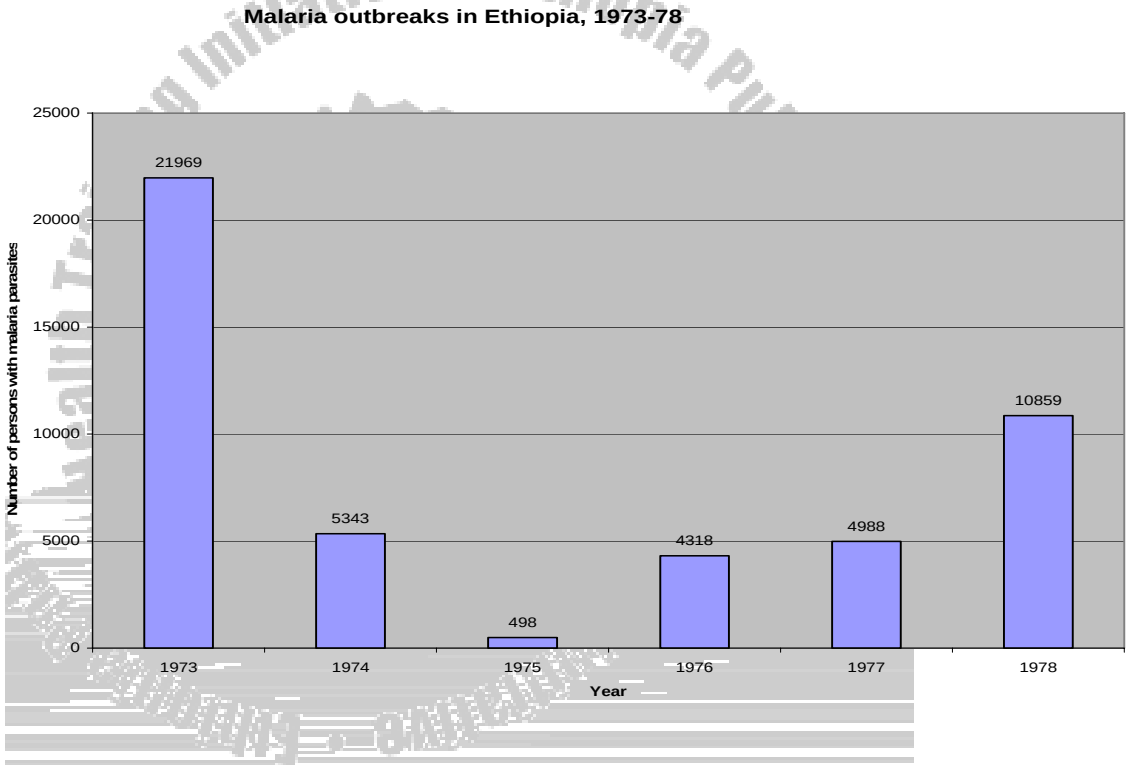


Key

Type of disease	Code numbers given
Intestinal parasites	1
Malaria	2
Skin diseases	3
Upper respiratory tract infections	4
Pneumonia	5
Gastritis	6
Diarrhoea	7
STI (sexually transmitted infections)	8
Eye diseases	9
Tuberculosis	10

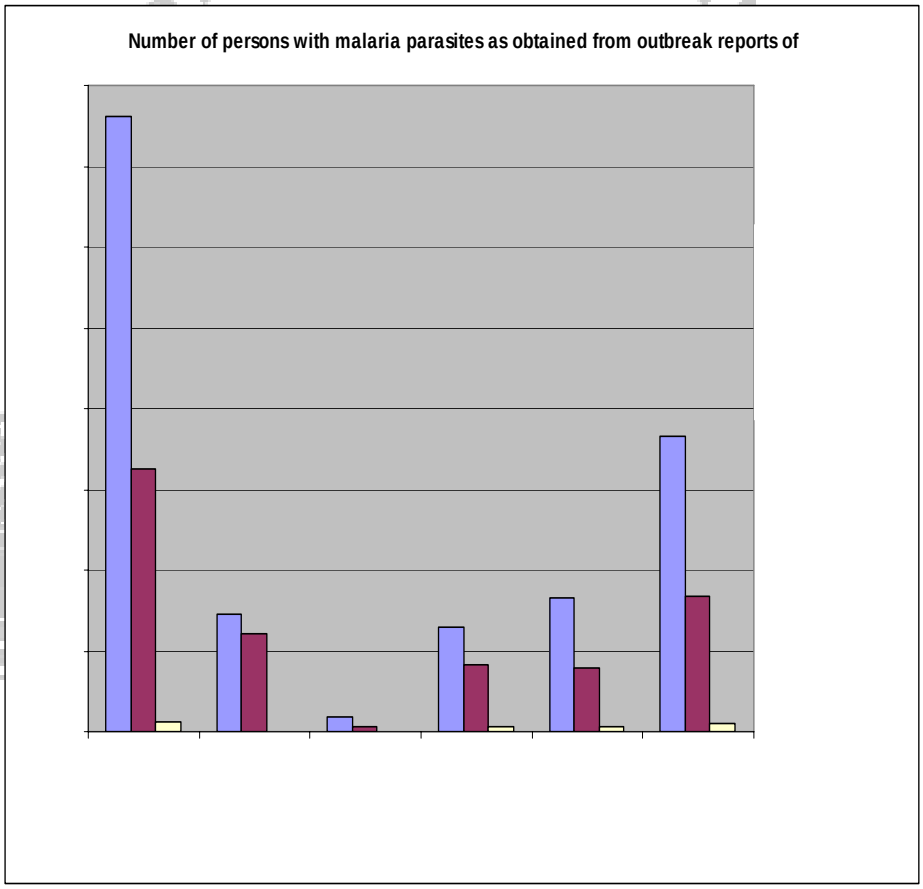
N.B. You can also write the type of disease in place of the number if enough space is available.

Example 2 : A simple bar chart showing malaria outbreaks in Ethiopia, 1973-78.



b) Multiple bar chart*:* In this type of chart the component figures are shown as separate bars adjoining each other. The height of each bar represents the actual value of the component figure. It depicts distributional pattern of more than one variable

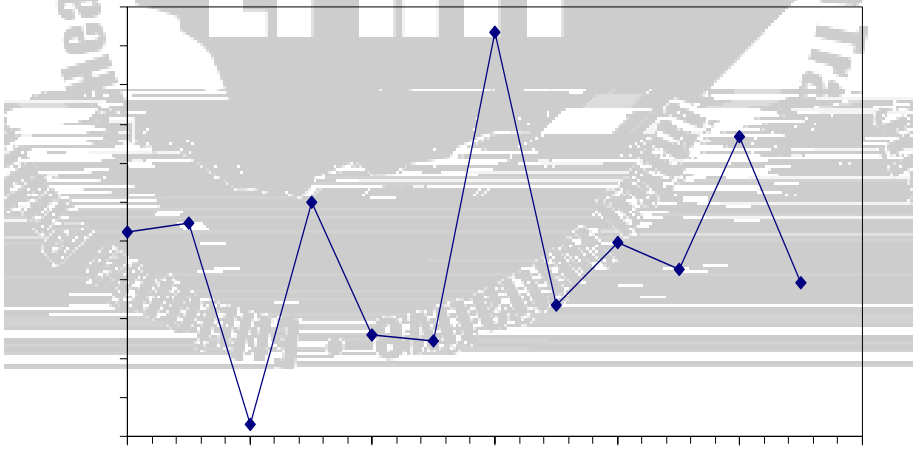
Example



2. The line graph

The line graph is especially useful for the study of some variables according to the passage of time. The time, in weeks, months or years is marked along the horizontal axis; and the value of the quantity that is being studied is marked on the vertical axis. The distance of each plotted point above the base-line indicates its numerical value and these points are joined by a line. The line graph is suitable for depicting a consecutive trend of a series over a long period.

Example: Malaria situation of Ethiopia as obtained from malaria seasonal blood survey results,1967-79 E.C., classified by slide positivity rate and year.



Health Training Initiative • Ethiopia Public Health Tra
EPHTI

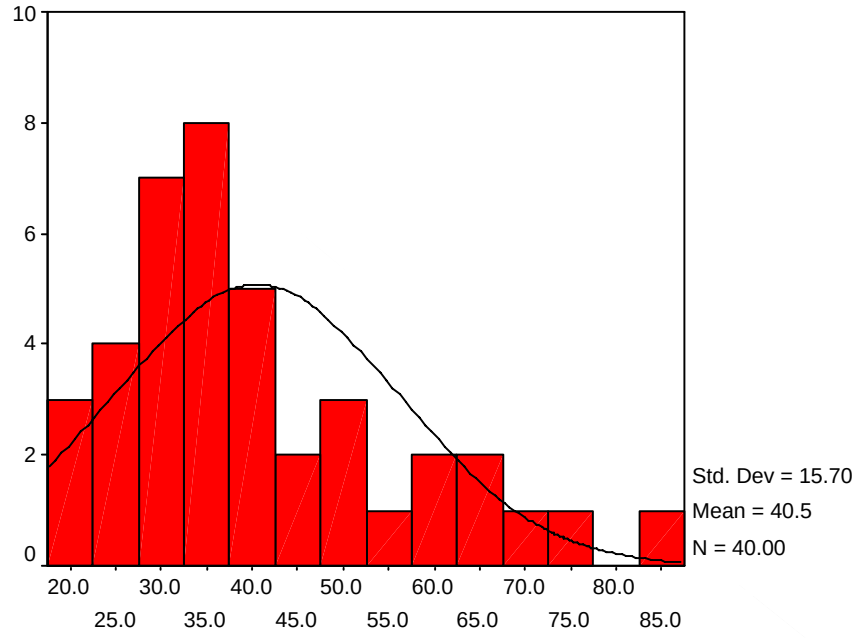
Example 1: Age of 40 persons who attended a meeting on one of the health days, Gondar, December 1995 (eth. cal):

20	30	30	27
31	33	55	29
32	38	33	29
49	46	35	59
49	42	23	75
84	29	35	58
49	35	40	64
21	25	24	70
22	35	40	67
47	33	29	35

When the above figures are rearranged in ascending order:

20	21	22	23	24	25	27	29	29	29
30	30	31	32	33	33	33	35	35	35
35	35	38	40	40	42	42	46	47	49
49	49	55	58	59	64	67	70	75	84

The age distribution of individuals who attended the health meeting on one of the health days, Gondar, 1995.



age of individuals who attended the health meeting

Mean	40.5 years
Median	35 years

The median is the middle score, when scores are arranged in increasing or decreasing order.

With 40 scores, the median is the mean of the 20th and 21st scores.

(see unit 5 for an exhaustive explanation)

Exercises

1. Explain the importance of tabular and graphical presentation of data.
2. The choice of a particular graph mostly depends on _____.
3. The following abstract (summary result) was taken from the annual report of Malaria and other Vector-borne diseases Control Organization (MVDCO).

The Malaria Seasonal Blood survey (SBS) results of 1979 (Eth. C.) which covered all regions of Ethiopia are stated in the following manner. Arrange them in a table so that the various parts of the table can be visualized easily.

a) In areas exempted from spray (0 spray round) the result of the survey showed 867, 313, 3 and 3 P.f., P.v., P.m. and mixed infection respectively.
b) In areas of one spray round a year, the same survey showed 1003, 172, 3 and 2 P.f., P.v., P.m. and mixed infection respectively.
c) In areas of two rounds of spray a year, 706 (P.f.), 200(P.v.), 4(P.m.) and 3(mixed infection) were observed.

4. The table below shows the populations and occurrence of deaths in 5 adjacent villages in 1987 (Eth. C.).

Village	Population Size	No. of deaths
A	10,000	200
B	6,000	180
C	8,000	224
D	2,000	100
E	5,000	175
Total	31,000	879

a) Identify the type of graph which is appropriate to present the above data.

b) Calculate the death rates and draw graph using these rates. What do you conclude from your graph ?

Death rate = $\frac{numberofdeaths}{populationsize} X 1000$

Health Training Initiative • Ethiopia Public Health Training
EPHTI

Composition (Structure): is the distribution of a population into its various groupings mainly by age and sex.

Example: The age and sex distribution of the Ethiopian population refers to the number (%) of the population falling in each age group (at each age) by sex.

Change: refers to the increase or decline of the total population or its components. The components of change are **birth, death, and migration**.

4.3 Sources of Demographic Data

Demographic information is acquired through two main ways: by complete enumerations (census) and sample surveys at a point in time, and through recording vital events as they occur over a period of time.

Complete enumerations or censuses are taken by obtaining information concerning every inhabitant of the area. Also coming into increasing use are sample surveys, conducted by interviewing a part of the population to represent the whole. On the other hand the information obtained from the recording of vital events (birth, death, marriage, divorce, etc) on a continuous basis completes the data collected from periodic censuses & sample surveys.

a) The Census

In modern usage, the term “census” refers to a nation-wide counting of population. It is obtained by interviewing each person or household. A census is a large and complicated undertaking. There are two main different schemes for enumerating a population in a census.

De jure: The enumeration (or count) is done according to the usual or legal place of residence.

De facto: The enumeration is done according to the actual place of residence on the day of the census.

A de jure count of the members of a household excludes temporary residents and visitors, but includes permanent residents who are temporarily away. A de facto count includes temporary residents and visitors, but excludes permanent residents who happen to be away on the day of the census.

The de facto census (recording of individuals wherever they are found – whether their presence in that place be permanent or temporary) is favoured by Britain while the United States of America has traditionally used the de jure (permanent residence) schemes. The two censuses carried out in Ethiopia in 1984 and 1994 included both de jure and de facto types of counts.

Advantages and Disadvantages of the two schemes

de jure:

a) Advantage

- It yields information relatively unaffected by seasonal and other temporary movements of people (i.e, it gives a picture of the permanent population).

b) Disadvantages

- Some persons may be omitted from the count while some others may be counted twice.
- In some situations, it is difficult to be sure just which is a person's usual or legal residence. (In places where mobility is high and no fixed residence is indicated)
- Information collected regarding persons away from home is often incomplete or incorrect.

de facto:

advantage:

- offers less chance of double counting

disadvantage:

- Population figures may be inflated or deflated by tourists, travelling salesmen, and other transients.

For most practical purposes, various combinations or modifications of the two-schemes (i.e. de jure & de facto) are used depending upon national needs and the enumeration plan followed .

Information to be collected

Sex, age, marital status, educational status, economic characteristics, place of birth, language, fertility, mortality , citizenship (nationality), living conditions (e.g. house-ownership, type of housing and the like), religion, etc..

Essential features (characteristics) of a census

1) Separate enumeration and recording of the characteristics of each individual
2) It should refer to people inhabiting a well-defined territory
3) The population should be enumerated with respect to a well defined point in time
4) It should be taken at regular intervals (usually every ten years)
5) In most countries the personal data collected in a census are not used for other than statistical purposes.
6) The compilation and publication of data by geographic areas and by basic demographic variables is an integral part of a census.

In short, the main characteristics of a census could be

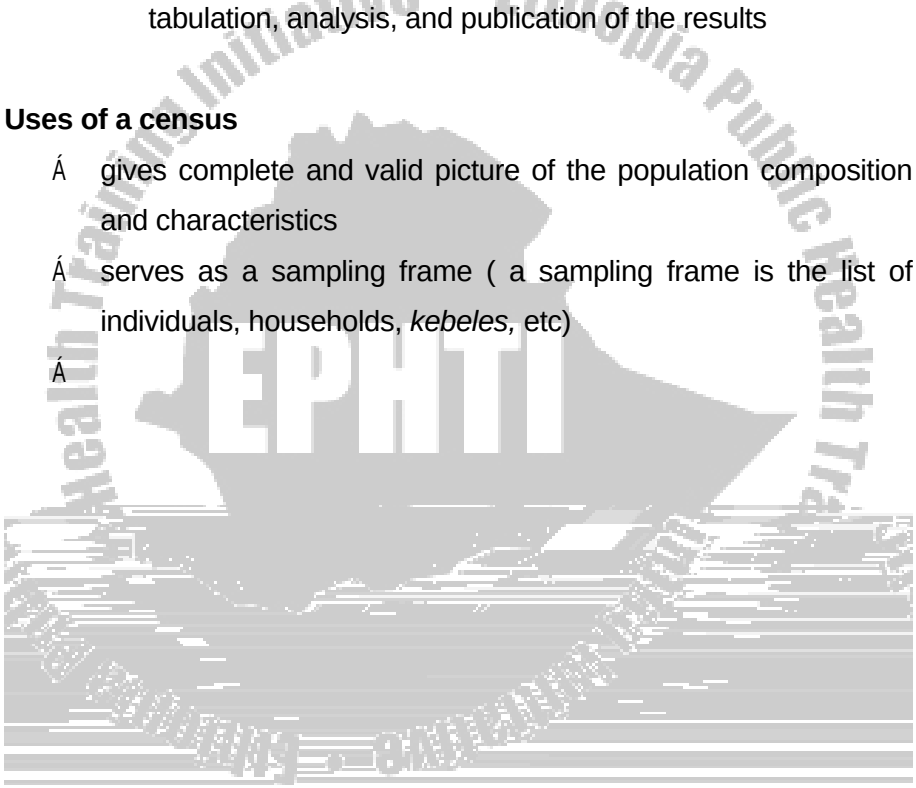
Census Operation

The entire census operation has 3 parts (stages)

1) pre-enumeration → planning and preparatory work
2) enumeration → field work (collection of the data)
3) post-enumeration → editing, coding, compilation, tabulation, analysis, and publication of the results

Uses of a census

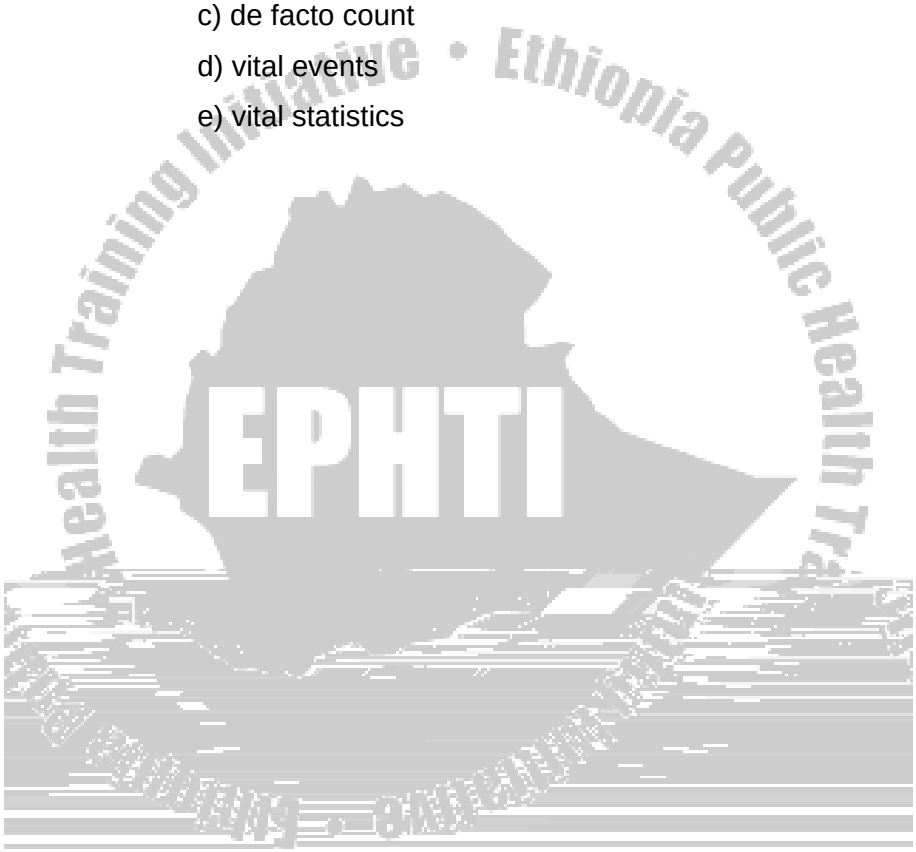
- gives complete and valid picture of the population composition and characteristics
- serves as a sampling frame (a sampling frame is the list of individuals, households, *kebeles,* etc)
-



c) Registration of vital events (mainly births and deaths)

Changes in population figures are taking place every day. Additions are made by births or through new arrivals from outside the area.

3. Explain the following terms by giving your own example for each of them.
   a) census
   b) de jure count
   c) de facto count
   d) vital events
   e) vital statistics



UNIT FIVE

Measures Of Central Tendency And Variation

5.1. Learning Objectives

At the end of this chapter, the student will be able to:

- Identify the different methods of data summarization
- Compute appropriate summary values for a set of data
- Appreciate the properties and limitations of summary values

5.2. Introduction

The first step in looking at data is to describe the data at hand in some concise way. In smaller studies this step can be accomplished by listing each data point. In general, however, this procedure is tedious or impossible and, even if it were possible would not give an over-all picture of what the data look like.

While tables are a convenient way to present specific information about individual values of a variable and graphs can provide a general picture of the pattern of the observations, it is often useful to provide, in addition, a numerical summary of the important characteristics of the distribution of a variable.

Before attempting the measures of central tendency and dispersion, let's see some of the notations that are used frequently.

Notations:

5.3. Measures of central tendency (location)

The tendency of statistical data to get concentrated at certain values is called the “Central Tendency” and the various methods of determining the actual value at which the data tend to concentrate are called measures of central Tendency or averages. Hence, an average is a value which tends to sum up or describe the mass of the data. For example, assume you have a huge mass of data. Naturally, most of the data values will concentrate at certain values and this tendency of concentration is called central tendency. The method of determining the actual value of the center of your data is called a measure of central tendency.

1. The Arithmetic Mean or simple Mean

One measure of central tendency is the arithmetic mean ; it is usually denoted by

3.0 1.9 4.0 3.6 2.9 3.5 3.0 2.8 3.2 2.9 2.2 3.0

What is the arithmetic mean for the sample birth weights?

The arithmetic mean is, in general, a very natural measure of central location. As a purely descriptive measure, however, the mean does have the disadvantage of being **seriously affected by extreme values**. In this instance it may not be representative of the location of the great majority of the sample points. For example, consider the following data regarding age at death (in years) of five persons from an organization. Fi

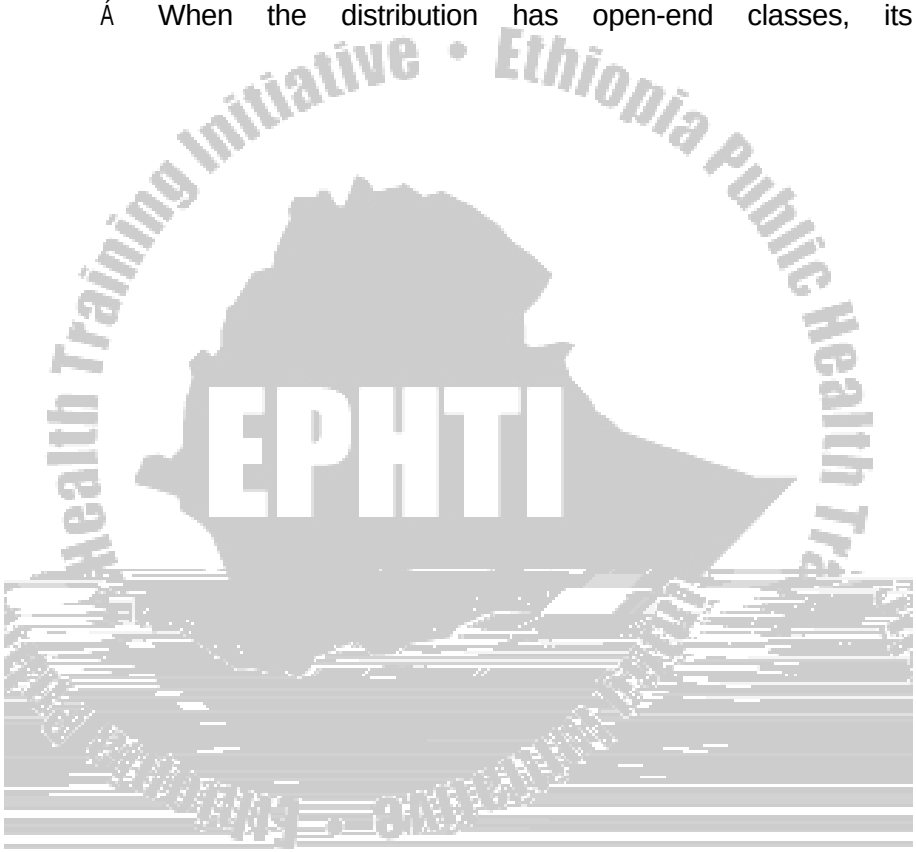
70 22 66 69 and 73

mean = (70+22+66+69+73) /5 = 60

Note how the mean is affected by the single observation of a death

b) Disadvantages

- It may be greatly affected by extreme items and its usefulness as a “summary of the whole” may be considerably reduced.
- When the distribution has open-end classes, its



Health Training Initiative • Ethiopia Public Health Training
EPHTI

Since n=12 is even,

Median = mean of the 6^{th} and 7^{th} largest observation = (3.0 + 3.0/2 = 3.0 kgs.

2) since n=5 which odd, we take the middle after arranging the observations in increasing or decreasing order.

22 66 69 70 73

Hence, the median is 69. In this case, the median is a better measure of central tendency.

The principal weakness of the median is that it is determined mainly by the middle points in a distribution and is less sensitive to the actual numerical values of the remaining data points.

a) Advantages

- It is easily calculated and is not much disturbed by extreme values.
- The median may be located even when the data are incomplete, e.g, when the class intervals are irregular and the final classes have open ends.

b) Disadvantages

- The median is not so well suited to algebraic treatment as the arithmetic and geometric means.
- It is not so generally familiar as the arithmetic mean.

3. Mode

It is the value of the observation that occurs with the greatest frequency. A particular disadvantage is that, with a small number of observations, there may be no mode. In addition, sometimes, there may be more than one mode such as when dealing with a bimodal (two-peaks) distribution. It is even less amenable (responsive) to mathematical treatment than the median.

Find the modal values for the following data (part 'a' is given in years while 'b' is given in kg)

a) 22, 66, 69, 70, 73 (no modal value)

b) 3.0, 1.9, 4.0, 3.6, 2.9, 3.5, 3.0, 2.8, 3.2, 2.9, 2.2, 3.0 (modal value = 3.0 kg)

a) Advantages

- Since it is the most typical value it is the most descriptive average
- Since the mode is usually an "actual value", it indicates the precise value of an important part of the series.

b) Disadvantages:-

- Unless the number of items is fairly large and the distribution reveals a distinct central tendency, the mode has no significance
- In a small number of items the mode may not exist.

Skewness: If extremely low or extremely high observations are present in a distribution, then the mean tends to shift towards those scores. Based on the type of skewness, distributions can be:

a) **Negatively skewed distribution**: occurs when majority of scores are at the right end of the curve and a few small scores are scattered at the left end.

Example: The ages of individuals who came to a health post seeking for a medical assistance are given below as follows (in increasing order).

15, 1 7, 41, 42, 43, 43, 44, 44, 45, 45, 46, 46, 46, 48, 50.

As can be seen from the above distribution, a few extreme **small** values (i.e., 15 and 17) are scattered at the left end and the mean will shift towards these scores. Therefore, the mean is not an appropriate measure of central tendency. You can see that the mean value is 41 which is less than the values of 12 of the observations. This shows that the mean is not a good summary measure of the central tendency of the above distribution. We say that this distribution is skewed to the left (i.e., negatively skewed). In this type

of distribution the median (which is 44) is a preferable measure of central tendency.

b) Positively skewed distribution: Occurs when the majority of scores are at the left end of the curve and a few extreme large scores are scattered at the right end.

Example: The ages of individuals who visited a health post on a given working day are given as follows (in increasing order):

19, 20, 20, 22, 22, 22, 23, 24, 25, 25, 26, 26, 27, 72, 77.

A few extreme **large** values (i.e., 72 and 77) are scattered at the right end and the mean will shift towards these scores. Therefore, the mean is not an appropriate measure of central tendency. You can see that the mean value is 30 which is greater than the values of 13 of the observations. This shows that the mean is not a good summary measure of the central tendency of the above distribution. We say that this distribution is skewed to the right (i.e., positively skewed). In this type of distribution the median is a preferable measure of central tendency. The median (middle value) is 24.

c) Symmetrical (Normal) distribution: It is neither positively nor negatively skewed. A curve is symmetrical if one half of the curve is the mirror image of the other half.

Example: The ages of individuals who visited a health post on one of the working days are given as follows (in increasing order):

19, 21, 22, 22, 23, 23, 24, 24, 24, 25, 25, 26, 26, 27, 29.

As can be seen from the above distribution, there are no extreme values like the ones we saw in parts "a" and "b". That is, the distribution is symmetrical. Therefore, the mean is an appropriate measure of central tendency. You can see that the mean value is 24 which is at the center of the distribution. This shows that the mean is a good summary measure of the central tendency of the above distribution. We say that this distribution is neither skewed to the left nor to the right (i.e., it is a symmetrical distribution).

As shown above, In a unimodal (one-peak) symmetrical distribution, the mean, median and mode are identical.

4. Geometric mean: It is obtained by taking the n^{th} root of the product of "n" values, i.e, if the values of the observations are demoted by $x_1, x_2, \ldots, x_n$ then, GM = $\sqrt[n]{(x1)(x2)\ldots(x_n)}$.

The geometric mean is preferable to the arithmetic mean if the series of observations contains one or more unusually large values. The above method of calculating geometric mean is satisfactory only if there are a small number of items. But if n is a large number, the problem of computing the n^{th} root of the product of these values by simple

arithmetic is a tedious work. To facilitate the computation of geometric mean we make use of logarithms. The above formula when reduced to its logarithmic form will be:

$GM = \sqrt[n]{(x1)(x2)\ldots(x_n)} = \{ (x1)(x2)\ldots (x_n) \}^{1/n}$

$\text{Log GM} = \log \{(x_1)(x_2)\ldots(x_n)\}^{1/n}$

$= 1/n \log \{(x_1)(x_2)\ldots(x_n)\}$

$= 1/n \{\log(x_1) + \log(x_2) + \ldots \log(x_n)\}$

$= \sum(\log xi)/n$

The logarithm of the geometric mean is equal to the arithmetic mean of the logarithms of individual values. The actual process involves obtaining logarithm of each value, adding them and dividing the sum by the number of observations. The quotient so obtained is then looked up in the tables of anti-logarithms which will give us the geometric mean.

Example: The geometric mean may be calculated for the following parasite counts per 100 fields of thick films.

7	8	3	14	2	1	440	15	52	6	2	1	1	25
12	6	9	2	1	6	7	3	4	70	20	200	2	50
21	15	10	120	8	4	70	3	1	103	20	90	1	237

GM = $\sqrt[42]{7x8x3x\ldots x90x1x237}$

log Gm = 1/42 (log 7+log8+log3+..+log 237)

= 1/42 (.8451+.9031+.4771 +…2.3747)

= 1/42 (41.9985)

= 0.9999 ≈ 1.0000

The anti-log of 0.9999 is 9.9992 ≈10 and this is the required geometric mean. By contrast, the arithmetic mean, which is inflated by the high values of 440, 237 and 200 is 39.8 ≈ 40.

a) Advantages

- since it is less affected by extreme values it is a more preferable average than the arithmetic mean.
- It is based on all values given in the distribution.

b) Disadvantages

- Its computation is relatively difficult.
- It cannot be determined if there is any negative value in the distribution, or where one of the items has a zero value.

5.4. Measures of Variation

In the preceding sections several measures which are used to describe the central tendency of a distribution were considered. While the mean, median, etc. give useful information about the center of the data, we also need to know how "spread out" the numbers are about the center.

Consider the following data sets:

Á

Health Training Initiative • Ethiopia Public Health Training
EPHTI

Example

Health Training Initiative • Ethiopia Public Health Training
EPHTI

3. Standard Deviation

Definition: The square root of the variance is called the standard deviation. The sample and population standard deviations denoted by S and σ (by convention) respectively are defined as follows:

S = $\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} = \sqrt{\text{sample variance}}$ = sample standard deviation

$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$ = population standard deviation (Note that , σ^2 = Population variance)

Example: Consider the above question (i.e., areas of sprayable surfaces with DDT from a sample of 15 houses).

101,105,110,114,115,124,125, 125, 130,133,135,136,137,140,145

Find the standard deviation of the above distribution.

The mean of the sample is 125 m^2.

Variance (sample) = $\mathbf{s^2} = \mathbf{\sum(Xi - \bar{X})^2/n-1}$

$= \{(101-125)^2 + (105-125)^2 + ...(145-125)^2\}/ (15-1)$

$= 2502/14$

$= 178.71$ (square metres)^2

Hence, the standard deviation = $\sqrt{178.71}$ = 13.37 m^2.

4. The coefficient of variation

The standard deviation is an absolute measure of deviation of observations around their mean and is expressed with the same unit of the data. Due to this nature of the standard deviation it is not directly used for comparison purposes with respect to variability. Therefore, it is useful to relate the **arithmetic mean ($\bar{X}$)** and **SD** (***S***) together, since, for example, a standard deviation of 10 would mean something different conceptually if the arithmetic mean were 10 than if it were 1000. A special measure called the coefficient of variation, is often used for this purpose.

Definition

The coefficient of variation is most useful in comparing the variability of several different samples, each with different means. This is because a higher variability is usually expected when the mean increases, and the CV is a measure that accounts for this variability. The coefficient of variation is also useful for comparing the reproducibility of different variables. CV is a relative measure free from unit of measurement. CV remains the same regardless of what units are used. In general, since CV is a pure number (a dimensionless quantity) it is suited for purposes of comparison.

Example: a) Compute the CV for the birth weight data when they are expressed in either grams or ounces.

Solution: in grams $\overline{X}$ = 3166.9 g, S = 445.3 g,

$$CV=100\% \times \frac{S}{\overline{X}} = 100\% \times \frac{445.3}{3166.9} = 14.1\%$$

If the data were expressed in ounces, $\overline{X}$ =111.71 oz, S=15.7 oz, then

CV = 100%

b)The following data were obtained from the summary results of 10 male medical students and the serum uric acid levels of 267 healthy males.

Type of variable	mean	Standard deviation
Pulse rate (beats/minute)	68.7	8.67
Serum uric acid level (mg per 100ml)	5.41	1.03

CV (pulse rate) = $\frac{8.67}{68.7}$ **x 100% = 12.6%**

CV (serum uric acid level) = $\frac{1.03}{5.41}$ **x 100 % = 19%**

Hence, the serum uric acid levels are relatively more spread out than pulse rates.

Exercises

1. Define arithmetic mean, median, mode and geometric mean and explain their strengths and limitations.
2. Define range, variance and standard deviation and explain their advantages and disadvantages.
3. When values of the three measures of central tendency

4. The mean ages of patients who visited a certain health facility from 1992 to 19994 are given as follows:

Year (Eth. C.)	1992	1993	1994
Mean age (in ears)	33	30	27
Number of patients	2000	4000	6000

What was the mean age of all the patients who visited the health facility from 1992 to 1994 ?

5. From the following data of calculation of arithmetic mean find the value of the missing figure.

Housing rent of each house in Birr	90.00	100.00	110.00	Missing figure
Number of houses	10	9	6	5

Total number of houses = 30, overall mean rent = Birr 99.00

6. Calculate the range and standard deviation of the above data (question no. 5)

References

1. Degu G. and Tessema F. Biostatistics for Health Science Students, lecture note series. Addis Ababa, January 2003.

2. Colton, T. Statistics in Medicine, 1st ed. ,Little, Brown and Company(inc), Boston, USA, 1974.

3. Bland, M. An Introduction to Medical Statistics, 3rd ed. University Press, Oxford, 2000.

4. Fletcher, M. Principles and Practice of Epidemiology, Addis Ababa; 1992.

5. Gupta C.B. An Introduction to Statistical Methods, 9th ed. Vikss publishing house PVT LTD, India; 1981.