LECTURE NOTES

For Health Science Students

Biostatistics

EPHTI
Ethiopia Public Health
Training Initiative

Getu Degu
Fasil Tessema

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

January 2005

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.

Important Guidelines for Printing and Photocopying
Limited permission is granted free of charge to print or photocopy all pages of this publication for educational, not-for-profit use by health care workers, students or faculty. All copies must retain all author credits and copyright notices included in the original document. Under no circumstances is it permissible to sell or distribute on a commercial basis, or to claim authorship of, copies of material reproduced from this publication.

©2005 by Getu Degu and Fasil Tessema

All rights reserved. Except as expressly provided above, no part of this publication may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system, without written permission of the author or authors.

This material is intended for educational use only by practicing health care workers or students and faculty in a health care field.

PREFACE

This lecture note is primarily for Health officer and Medical students who need to understand the principles of data collection, presentation, analysis and interpretation. It is also valuable to diplomastudents of environmental health, nursing and laboratory technology although some of the topics covered are beyond their requirements. The material could also be of paramount importance for an individual who is interested in medical or public health research.

It has been a usual practice for a health science student in Ethiopia to spend much of his/her time in search of reference materials on Biostatistics. Unfortunately, there are no textbooks which could appropriately fulfill the requirements of the Biostatistics course at the undergraduate level for Health officer and Medical students. We firmly believe that this lecture note will fill that gap.

The first three chapters cover basic concepts of Statistics focusing on the collection, presentation and summarization of data. Chapter four deals with the basic demographic methods and health service statistics giving greater emphasis to indices relating to the hospital. In chapters five and six elementary probability and sampling methods are presented with practical examples. A relatively comprehensive description of statistical inference on means and proportions is given in chapters seven and eight. The last chapter of this lecture note is about linear correlation and regression.

General learning objectives followed by introductory sections which are specific to each chapter are placed at the beginning of each chapter. The lecture note also includes many problems for the student, most of them based on real data, the majority with detailed solutions. A few reference materials are also given at the end of the lecture note for further reading.

Acknowledgments

We would like to thank the Gondar College of Medical Sciences and the Department of Epidemiology and Biostatistics (Jimma University) for allowing us to use the institutions resources while writing this lecture note. We are highly indebted to the Carter Center with out whose uninterrupted follow up and support this material would have not been written. we wish to thank our students whom we haveinstructed over the past years for their indirect contribution to the writing of this lecture note.

Table of Contents

Preface **i**
Acknowledgements **iii**
Table of contents **iv**

Chapter One : Introduction to Statistics

1.1 Learning Objectives 1
1.2 Introduction 1
1.3 Rationale of studying Statistics 5
1.4 Scales of measurement 7

Chapter Two: Methods of Data collection, Organization and presentation

2.1 Learning Objectives 12
2.2 Introduction 12
2.3 Data collection methods 13
2.4 Choosing a method of data collection 19
2.5 Types of questions 22
2.6 Steps in designing a questionnaire 27
2.7 Methods of data organization and presentation 32

Chapter Three : Summarizing data

3.1 Learning Objectives 61
3.2 Introduction 61
3.3 Measures of Central Tendency 63
3.4 Measures of Variation 74

Chapter Four : Demographic Methods and Health Services Statistics

4.1 Learning Objectives 95
4.2 Introduction 95
4.3 Sources of demographic data 97
4.4 Stages in demographic transition 103
4.5 Vital Statistics 107
4.6 Measures of Fertility 109
4.7 Measures of Mortality 114
4.8 Population growth and Projection 117
4.9 Health services statistics 119

Chapter Five : Elementary Probability and probability distribution

5.1 Learning Objectives 126
5.2 Introduction 126
5.3 Mutually exclusive events and the additive law 129

5.4 Conditional Probability and the multiplicative law 131
5.5 Random variables and probability distributions 135

Chapter Six : Sampling methods

6.1 Learning Objectives 150
6.2 Introduction 150
6.3 Common terms used in sampling 151
6.4 Sampling methods 153
6.5 Errors in Sampling 160

Chapter seven : Estimation

7.1 Learning Objectives 163
7.2 Introduction 163
7.3 Point estimation. 164
7.4 Sampling distribution of means 165
7.5 Interval estimation (large samples) 169
7.6 Sample size estimation 179
7.7 Exercises 185

Chapter Eight : Hypothesis Testing

8.1 Learning Objectives 186
8.2 Introduction 186
8.3 The Null and Alternative Hypotheses 188
8.4 Level of significance 191

8.5 Tests of significance on means and proportions (large samples) 193
8.6 One tailed tests 204
8.7 Comparing the means of small samples 208
8.8 Confidence interval or P-value? 219
8.9 Test of significance using the Chi-square and Fisher's exact tests 221
8.10 Exercises 229

Chapter Nine: Correlation and Regression

9.1 Learning Objectives 231
9.2 Introduction 231
9.3 Correlation analysis 232
9.4 Regression analysis 241

Appendix : Statistical tables 255

References 263

List of Tables

Table 1 overall immunization status of children in adamai Tullu Woreda, Feb 1995 46
Table 2: TT immunization by marital status of the woment of childbearing age, assendabo town jimma Zone, 1996 47
Table 3 Distribution of Health professional by sex and residence 48
Table 4 Area in one tall of the standard normal curve 255
Table 5 Percentage points of the t Distribution 258
Table 6 Percentage points of the Chi-square distribution 260

List of Figures

Figure 1 Immunization status of children in Adami Tulu woreda, Feb. 1995 52
Figure 2 TT Immunization status by marital status of women 15-49 year, Asendabo town, 1996 53
Figure 3 TT immunization status by marital status of women 15-49 years, Asendabo town, 1996TT 54
Figure 4 Immunization status by marital status of women 15-49 years, Asendabo town 1996 55
Figure 5 Immunization status of children in Adami Tulu woreda, Feb. 1995 55
Figure 6 Histogram for amount of time college students devoted to leisure activities 57
Figure 7 Frequency polygon curve on time spent for leisure activities by students Cumulative 58
Figure 8 frequency curve for amountof time college students devoted to leisure activities 59
Figure 9 Malaria parasite rates in Ethiopia, 1967-1979Eth. c. 60

CHAPTER ONE

Introduction to Statistics

1.1. Learning objectives

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

1. Define Statistics and Biostatistics
2. Enumerate the importance and limitations of statistics
3. Define and Identify the different types of data and understand why we need to classifying variables

1.2. Introduction

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

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 malaria detection and treatment 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:

i) They must be in aggregates – This means that statistics are 'number of facts.' A single fact, even though numerically stated, cannot be called statistics.

ii) 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 outbreak 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 to assess the individual contribution of any one of these factors.

iii) They must be enumerated or estimated according to a reasonable standard of accuracy – Statistics must be enumerated or estimated according to reasonable standards 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.

iv) 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 planned manner and for a purpose about which the enumerator had a definite idea. Facts collected in an unsystematic manner and without a complete awareness of the object, will be confusing and cannot be made the basis of valid conclusions.

v) They must be placed in relation to each other. That is, they must be comparable. Numerical facts may be placed in relation to each other either in point of time, space or condition. The phrase, ‘placed in relation to each other' suggests that the facts should be comparable.

Also included in this view are the techniques for tabular and graphical presentation of data as well as the methods used to summarize a body of data with one or two meaningful figures. This aspect of organization, presentation and summarization of data are labelled as ***descriptive statistics***.

One branch of descriptive statistics of special relevance in medicine is that of vital statistics – vital events: birth, death, marriage, divorce, and the occurrence of particular disease. 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 that pervade both medical and lay literature.

statistical methods: When the term 'statistics' is used to mean 'statistical methods' it refers to a body of methods that are used for collecting, organising, analyzing and interpreting numerical data for understanding a phenomenon or making wise decisions. In this sense it is a branch of scientific method and helps us to know in a better way the object under study.

The branch of modern statistics that is most relevant to public health and clinical medicine is ***statistical inference.*** This branch of statistics deals with techniques of making conclusions about the population. Inferential statistics builds upon descriptive statistics. The inferences are drawn from particular properties of sample to particular properties of population. These are the types of statistics most commonly found in research publications.

Definition: When the different statistical methods are applied in biological, medical and public health data they constitute the discipline of Biostatistics..

1.3 Rationale of studying statistics

- Statistics pervades a way of organizing information on a wider and more formal basis than relying on the exchange of anecdotes and personal experience
- More and more things are now measured quantitatively in medicine and public health
- There is a great deal of intrinsic (inherent) variation in most biological processes
- Public health and medicine are becoming increasingly quantitative. As technology progresses, the physician encounters more and more quantitative rather than descriptive information. Inone sense, statistics is the language of assembling and handling quantitative material. Even if one's concern is only with the results of other people's manipulation and assemblage of data, itis important to achieve some understanding of this language in order to interpret their results properly.
- The planning, conduct, and interpretation of much of medical research are becoming increasingly reliant on statistical technology. Is this new drug or procedure better than the one commonly in use? How much better? What, if any, are the risks of side effects associated with its use? In testing a new drug how many patients must be treated, and in what manner, in order to demonstrate its worth? What is the normal variation in some clinical measurement? How reliable and valid is the

measurement? What is the magnitude and effect of laboratory and technical error? How does one interpret abnormal values?

- Statistics pervades the medical literature. As a consequence of the increasingly quantitative nature of public health and medicine and its reliance on statistical methodology, the medical literatureis replete with reports in which statistical techniques are used extensively.

"It is the interpretation of data in the presence of such variability that lays at the heart of statistics."

Limitations of statistics:

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

1. It deals on aggregates of facts and no importance is attached to individual items–suited only if their group characteristics are desired to be studied.
2. Statistical data are only approximately and not mathematically correct.

1.4 Scales of measurement

Any aspect of an individual that is measured and take any value for different individuals or cases, like blood pressure, or records, like age, sex is called a ***variable.***

It is helpful to divide variables into different types, as different statistical methods are applicable to each. The main division is into qualitative (or categorical) or quantitative (or numerical variables).

Qualitative variable: a variable or characteristic which cannot be measured in quantitative form but can only be identified by name or categories, for instance place of birth, ethnic group, type of drug, stages of breast cancer (I, II, III, or IV), degree of pain (minimal, moderate, severe or unbearable).

Quantitative variable: A quantitative variable is one that can be measured and expressed numerically and they can be of two types (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.

The simplest data consist of unordered, dichotomous, or "either - or" types of observations, i.e., either the patient lives or the patient dies, either he has some particular attribute or he does not.

eg. Nominal scale data: survival status of propanolol - treated and control patients with myocardial infarction

Status 28 days after hospital admission	Propranolol -treated patient	Control Patients
Dead	7	17
Alive	38	29
Total Survival rate	45 84%	46 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 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 survived the first 28 days after hospital admission or he succumbed (died) sometime within this time period.

With nominal scale data the obvious and intuitive descriptive summary measure is the proportion or percentage of subjects who exhibit the attribute. Thus, we can see from the above table that 84 percent of the patients treated with propranolol survived, in contrast with only 63% of the control group.

Some other examples of nominal data:

Eye color - brown, black, etc.
Religion - Christianity, Islam, Hinduism, etc
Sex - male, female

Ordinal Data:- have order among the response classifications (categories). The spaces or intervals between the categories are not necessarily equal.

Example:

1. strongly agree
2. agree
3. no opinion
4. disagree
5. strongly disagree

In the above situation, we only know that the data are ordered.

Interval Data:- In interval data the intervals between values are the same. For example, in the Fahrenheit temperature scale, the difference between 70 degrees and 71 degrees is the same as the difference between 32 and 33 degrees. But the scale is not a RATIO Scale. 40 degrees Fahrenheit is not twice as much as 20 degrees Fahrenheit.

Ratio Data:- The data values in ratio data do have meaningful ratios, for example, age is a ratio data, some one who is 40 is twice as old as someone who is 20.

Both interval and ratio data involve measurement. Most data analysis techniques that apply to ratio data also apply to interval data. Therefore, in most practical aspects, these types of data (interval and ratio) are grouped under metric data. In some other instances, these type of data are also known as numerical discrete and numerical continuous.

Numerical discrete

Numerical discrete data occur when the observations are integers that correspond with a count of some sort. Some common examples are: the number of bacteria colonies on a plate, the number of cells within a prescribed area upon microscopic examination, the number of heart beats within a specified time interval, a mother's history of number of births (parity) and pregnancies (gravidity), the number of episodes of illness a patient experiences during some time period, etc.

Numerical continuous

The scale with the greatest degree of quantification is a numerical continuous scale. Each observation theoretically falls somewhere along a continuum. One is not restricted, in principle, to particular values such as the integers of the discrete scale. The restricting factor is the degree of accuracy of the measuring instrument most clinical measurements, such as blood pressure, serum cholesterol level, height, weight, age etc. are on a numerical continuous scale.

1.5 Exercises

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

1. Blood group
2. Temperature (Celsius)
3. Ethnic group
4. Job satisfaction index (1-5)
5. Number of heart attacks
6. Calendar year
7. Serum uric acid (mg/100ml)
8. Number of accidents in 3 - year period
9. Number of cases of each reportable disease reported by a health worker
10. The average weight gain of 6 1-year old dogs (with a special diet supplement) was 950grams last month.

CHAPTER TWO

Methods Of Data Collection, Organization And Presentation

2.1. Learning Objectives

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

1. Identify the different methods of data organization and presentation
2. Understand the criterion for the selection of a method to organize and present data
3. Identify the different methods of data collection and criterion that we use to select a method of data collection
4. Define a questionnaire, identify the different parts of a questionnaire and indicate the procedures to prepare a questionnaire

2.2. Introduction

Before any statistical work can be done data must be collected. Depending on the type of variable and the objective of the study different data collection methods can be employed.

2.3. Data Collection Methods

Data collection techniques allow us to systematically collect dataabout 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 research questions in a conclusive way.

Various data collection techniques can be used such as:

- Observation
- Face-to-face and self-administered interviews
- Postal or mail method and telephone interviews
- Using available information
- Focus group discussions (FGD)
- Other data collection techniques – Rapid appraisal techniques, 3L technique, Nominal group techniques, Delphi techniques, life histories, case studies, etc.

***1. Observation* –** Observation is a technique that involves systematically selecting, watching and recoding behaviors of people or other phenomena and aspects of the setting in which they occur, for the purpose of getting (gaining) specified information. It includes all methods from simple visual observations to the use of high level machines and measurements, sophisticated equipment or facilities, such as radiographic, biochemical, X-ray machines, microscope, clinical examinations, and microbiological examinations.

Outline the guidelines for the observations prior to actual data collection.

Advantages: Gives relatively more accurate data on behavior 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. Interviews and self-administered questionnaire

Interviews and self-administered questionnaires are probably the most commonly used research data collection techniques. Therefore, designing good "***questioning tools***" forms an important and time consuming phase in the development of most research proposals.

Once the decision has been made to use these techniques, the following questions should be considered before designing our tools:

- What exactly do we want to know, according to the objectives and variables we identified earlier? Is questioning the right technique to obtain all answers, or do we need additional techniques, such as observations or analysis of records?

- Of whom will we ask questions and what techniques will we use? Do we understand the topic sufficiently to design a questionnaire, or do we need some loosely structured interviews with key informants or a focus group discussion first to orient ourselves?

- Are our informants mainly literate or illiterate? If illiterate, the use of self-administered questionnaires is not an option.
- How large is the sample that will be interviewed? Studies with many respondents often use shorter, highly structured questionnaires, whereas smaller studies allow more flexibility and may use questionnaires with a number of open-ended questions.

Once the decision has been made Interviews may be less or more structured. Unstructured interview is flexible, the content wording and order of the questions vary from interview to interview. The investigators only have idea of what they want to learn but do not decide in advance exactly what questions will be asked, or in what order.

In other situations, a more standardized technique may be used, the wording and order of the questions being decided in advance. This may take the form of a highly structured interview, in which the questions are asked orderly, or a self administered questionnaire, in which case the respondent reads the questions and fill in the answers by himself (sometimes in the presence of an interviewer who ‘stands by’ to give assistance if necessary).

Standardized methods of asking questions are usually preferred in community medicine research, since they provide more assurance that the data will be reproducible. Less structured interviews may be useful in a preliminary survey, where the purpose is to obtain information to help in the subsequent planning of a study rather than factors for analysis, and in intensive studies of perceptions, attitudes, motivation and affective reactions. Unstructured interviews are characteristic of qualitative (non-quantitative) research.

The use of self-administered questionnaires is simpler and cheaper; such questionnaires can be administered to many persons simultaneously (e.g. to a class of students), and unlike interviews, can be sent by post. On the other hand, they demand a certain level of education and skill on the part of the respondents; people of a low socio-economic status are less likely to respond to a mailed questionnaire.

In interviewing using questionnaire, the investigator appoints agents known as enumerators, who go to the respondents personally with the questionnaire, ask them the questions given there in, and record their replies. They can be either face-to-face or telephone interviews.

Face-to-face and telephone interviews have many advantages. A good interviewer can stimulate and maintain the respondent's interest, and can create a rapport (understanding, concord) and atmosphere conducive to the answering of questions. If anxiety aroused, the interviewer can allay it. If a question is not understood an interviewer can repeat it and if necessary (and in accordance with guidelines decided in advance) provide an explanation or alternative wording. Optional follow-up or probing questions that are to be askedonly if prior responses are inconclusive or inconsistent cannot easily be built into self-administered questionnaires. In face-to-face interviews, observations can be made as well.

In general, apart from their expenses, interviews are preferable to self-administered questionnaire, with the important proviso that they are conducted by skilled interviewers.

Mailed Questionnaire Method: Under this method, the investigator prepares a questionnaire containing a number of questions pertaining the field of inquiry. The questionnaires are sent by post to the informants together with a polite covering letter explaining the detail, the aims and objectives of collecting the information, and requesting the respondents to cooperate by furnishing the correct replies and returning the questionnaire duly filled in. In order to ensure quick response, the return postage expenses are usually borne by the investigator.

The main problems with postal questionnaire are that response rates tend to be relatively low, and that there may be under representation of less literate subjects.

3. Use of documentary sources: Clinical and other personal records, death certificates, published mortality statistics, census publications, etc. Examples include:

1. Official publications of Central Statistical Authority
2. Publication of Ministry of Health and Other Ministries
3. News Papers and Journals.
4. International Publications like Publications by WHO, World Bank, UNICEF
5. Records of hospitals or any Health Institutions.

During the use of data from documents, though they are less time consuming and relatively have low cost, care should be taken on the quality and completeness of the data. There could be differences in objectives between the primary author of the data and the user.

Problems in gathering data

It is important to recognize some of the main problems that may be faced when collecting data so that they can be addressed in the selection of appropriate collection methods and in the training of the staff involved.

Common problems might include:

- Language barriers
- Lack of adequate time
- Expense
- Inadequately trained and experienced staff
- Invasion of privacy
- Suspicion
- Bias (spatial, project, person, season, diplomatic, professional)
- Cultural norms (e.g. which may preclude men interviewing women)

2.4. Choosing a Method of Data Collection

Decision-makers need information that is ***relevant, timely, accurate and usable***. The cost of obtaining, processing and analyzing these data is high. The challenge is to find ways, which lead to information that is cost-effective, relevant, timely and important for immediate use. Some methods pay attention to ***timeliness and reduction in cost***. Others pay attention to ***accuracy and the strength of the method*** in using scientific approaches.

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

1) Primary data 2) Secondary data

Primary Data: are those data, which are collected by the investigator himself 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 hisown purposes.

The secondary data can be obtained from journals, reports, government publications, publications of professionals and research organizations.

Secondary data are less expensive to collect both in money and time. These data can also be better utilized and sometimes the quality of such data may be better because these might have been collected by persons who were specially trained for that purpose.

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. Secondly, 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.

In general, the choice of methods of data collection is largely based on the ***accuracy of the information they yield***. In this context, 'accuracy' refers not only to correspondence between the information and objective reality - although this certainly enters into the concept - but also to the information's relevance. This issue is the extent to which the method will provide a precise measure of the variable the investigator wishes to study.

The selection of the method of data collection is also based on practical considerations, such as:

1) The need for personnel, skills, equipment, etc. in relation to what is available and the urgency with which results are needed.

2) The acceptability of the procedures to the subjects - the absence of inconvenience, unpleasantness, or untoward consequences.
3) The probability that the method will provide a good coverage, i.e. will supply the required information about all or almost all members of the population or sample. If many people will not know the answer to the question, the question is not an appropriate one.

The investigator's familiarity with a study procedure may be a valid consideration. It comes as no particular surprise to discover that a scientist formulates problems in a way which requires for their solution just those techniques in which he himself is specially skilled.

2.5. Types of Questions

Before examining the steps in designing a questionnaire, we need to review the types of questions used in questionnaires. Depending on how questions are asked and recorded we can distinguish two major possibilities - Open –ended questions, and closed questions.

Open-ended questions

Open-ended questions permit free responses that should be recorded in the respondent's own words. The respondent is not given any possible answers to choose from.

Such questions are useful to obtain information on:

- Facts with which the researcher is not very familiar,
- Opinions, attitudes, and suggestions of informants, or
- Sensitive issues.

For example

“Can you describe exactly what the traditional birth attendant did when your labor started?”

“What do you think are the reasons for a high drop-out rate of village health committee members?”

“What would you do if you noticed that your daughter (school girl) had a relationship with a teacher?”

Closed Questions

Closed questions offer a list of possible options or answers from which the respondents must choose. When designing closed questions one should try to:

- Offer a list of options that are exhaustive and mutually exclusive
- Keep the number of options as few as possible.

Closed questions are useful if the range of possible responses is known.

For example

“What is your marital status?

1. Single ○
2. Married/living together ○
3. Separated/divorced/widowed ○

“Have your every gone to the local village health worker for treatment?

1. Yes ○
2. No ○

Closed questions may also be used if one is only interested in certain aspects of an issue and does not want to waste the time of the respondent and interviewer by obtaining more information than one needs.

For example, a researcher who is only interested in the proteincontent of a family diet may ask:

“Did you eat any of the following foods yesterday? (Circle yes or no for each set of items)

▪ Peas, bean, lentils	Yes	No
▪ Fish or meat	Yes	No
▪ Eggs	Yes	No
▪ Milk or Cheese	Yes	No

Closed questions may be used as well to get the respondents to express their opinions by choosing rating points on a scale.

For example

“How useful would you say the activities of the Village Health Committee have been in the development of this village?”

1. Extremely useful ○
2. Very useful ○
3. Useful ○
4. Not very useful ○
5. Not useful at all ○

Requirements of questions

Must have face validity – that is the question that we design should be one that give an obviously valid and relevant measurement for the variable. For example, it may be self-evident that records kept in an obstetrics ward will provide a more valid indication of birth weights than information obtained by questioning mothers.

Must be clear and unambiguous – the way in which questions are worded can ‘*make or break*’ a questionnaire. Questions must be clear and unambiguous. They must be phrased in language that it is believed the respondent will understand, and that all respondents will understand in the same way. To ensure clarity, each question should contain only one idea; ‘double-barrelled’ questions like ‘Do you take your child to a doctor when he has a cold or has diarrhoea?’ are difficult to answer, and the answers are difficult to interpret.

Must not be offensive – whenever possible it is wise to avoid questions that may offend the respondent, for example those thatdeal with intimate matters, those which may seem to expose the respondent’s ignorance, and those requiring him to give a socially unacceptable answer.

The questions should be fair - They should not be phrased in a way that suggests a specific answer, and should not be loaded. Short questions are generally regarded as preferable to long ones.

Sensitive questions - It may not be possible to avoid asking ‘sensitive’ questions that may offend respondents, e.g. those that seem to expose the respondent’s ignorance. In such situations the interviewer (questioner) should do it very carefully and wisely

2.6 Steps in Designing a Questionnaire

Designing a good questionnaire always takes several drafts. In the first draft we should concentrate on the content. In the second, we should look critically at the formulation and sequencing of the questions. Then we should scrutinize the format of the questionnaire.Finally, we should do a test-run to check whether the questionnaire gives us the information we require and whether both the respondents and we feel at ease with it. Usually the questionnaire will need some further adaptation before we can use it for actual data collection.

Step1: CONTENT

Take your objectives and variables as your starting point.

Decide what questions will be needed to measure or to define your variables and reach your objectives. When developing the questionnaire, you should reconsider the variables you have chosen, and, if necessary, add, drop or change some. You may even change some of your objectives at this stage.

Step 2: FORMULATING QUESTIONS

Formulate one or more questions that will provide the information needed for each variable.

Take care that questions are specific and precise enough that different respondents do not interpret them differently. For example, a question such as: “Where do community members usually seek treatment when they are sick?” cannot be asked in such a generalway because each respondent may have something different in mind when answering the question:

- One informant may think of measles with complications and say he goes to the hospital, another of cough and say goes to the private pharmacy;

- Even if both think of the same disease, they may have different degrees of seriousness in mind and thus answerdifferently;
- In all cases, self-care may be overlooked.

The question, therefore, as rule has to be broken up into different parts and made so specific that all informants focus on the samething. For example, one could:

- Concentrate on illness that has occurred in the family over the past 14 days and ask what has been done to treat if from the onset; or
- Concentrate on a number of diseases, ask whether they have occurred in the family over the past X months (chronic or serious diseases have a longer recall period than minor ailments) and what has been done to treat each of them from the onset.

Check whether each question measures one thing at a time.

For example, the question, "How large an interval would you and your husband prefer between two successive births?" would better be divided into two questions because husband and wife may have different opinions on the preferred interval.

Avoid leading questions.

A question is leading if it suggests a certain answer. For example, the question, "Do you agree that the district health team should visit each health center monthly?" hardly leaves room for “no” or for other options. Better would be: “Do you thing that district health teams should visit each health center? If yes, how often?”

Sometimes, a question is leading because it presupposes a certain condition. For example: “What action did you take when your child had diarrhoea the last time?” presupposes the child has had diarrhoea. A better set of questions would be: “Has your child had diarrhoea? If yes, when was the last time?” “Did you do anything to treat it? If yes, what?”

Step 3: SEQUENCING OF QUESTIONS

Design your interview schedule or questionnaire to be “consumer friendly.”

- The sequence of questions must be logical for the respondent and allow as much as possible for a “natural” discussion, even in more structured interviews.
- At the beginning of the interview, keep questions concerning “background variables” (e.g., age, religion, education, marital status, or occupation) to a minimum. If possible, pose most or all of these questions later in the interview. (Respondents

may be reluctant to provide “personal” information early in an interview)

- Start with an interesting but non-controversial question (preferably open) that is directly related to the subject of the study. This type of beginning should help to raise the informants’ interest and lessen suspicions concerning the purpose of the interview (e.g., that it will be used to provide information to use in levying taxes).
- Pose more sensitive questions as late as possible in the interview (e.g., questions pertaining to income, sexual behavior, or diseases with stigma attached to them, etc.
- Use simple everyday language.

Make the questionnaire as short as possible. Conduct the interview in two parts if the nature of the topic requires a long questionnaire (more than 1 hour).

Step 4: FORMATTING THE QUESTIONNAIRE

When you finalize your questionnaire, be sure that:

- Each questionnaire has a heading and space to insert the number, data and location of the interview, and, if required the name of the informant. You may add the name of the interviewer to facilitate quality control.

- Layout is such that questions belonging together appear together visually. If the questionnaire is long, you may use subheadings for groups of questions.
- Sufficient space is provided for answers to open-ended questions.
- Boxes for pre-categorized answers are placed in a consistent manner half of the page.

Your questionnaire should not only be consumer but also user friendly!

Step 5: TRANSLATION

If interview will be conducted in one or more local languages, the questionnaire has to be translated to standardize the way questions will be asked. After having it translated you should have itretranslated into the original language. You can then compare the two versions for differences and make a decision concerning the final phrasing of difficult concepts.

2.7 Methods of data organization and presentation

The data collected in a survey is called ***raw data***. In most cases, useful information is not immediately evident from the mass of unsorted 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. Precise methods of analysis can be decided up on only when the characteristics of the data are understood. For the primary objective of this different techniques of data organization and presentation like order array, tables and diagrams are used.

2.7.1 Frequency Distributions

For data to be more easily appreciated and to draw quick comparisons, it is often useful to arrange the data in the form of a table, or in one of a number of different graphical forms.

When analysing voluminous data collected from say, a health center's records, it is quite useful to put them into compact tables. Quite often, the presentation of data in a meaningful way is done by preparing a frequency distribution. If this is not done the raw data will not present any meaning and any pattern in them (if any) may not be detected.

Array (ordered array) is a serial arrangement of numerical data in an ascending or descending order. This will enable us to know the range over which the items are spread and will also get an idea of their general distribution. Ordered array is an appropriate way of presentation when the data are small in size (usually less than 20).

A study in which 400 persons were asked how many full-length movies they had seen on television during the preceding week. The following gives the distribution of the data collected.

Number of movies	Number of persons	Relative frequency (%)
0	72	18.0
1	106	26.5
2	153	38.3
3	40	10.0
4	18	4.5
5	7	1.8
6	3	0.8
7	0	0.0
8	1	0.3
Total	400	100.0

In the above distribution Number of movies represents the variable under consideration, Number of persons represents the frequency, and the whole distribution is called frequency distribution particularly simple frequency distribution.

A categorical distribution – non-numerical information can also be represented in a frequency distribution. Seniors of a high schoolwere interviewed on their plan after completing high school. The following data give plans of 548 seniors of a high school.

SENIORS' PLAN	NUMBER OF SENIORS
Plan to attend college	240
May attend college	146
Plan to or may attend a vocational school	57
Will not attend any school	105
Total	548

Consider the problem of a social scientist who wants to study the age of persons arrested 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
Under 18	1,748
18 – 24	3,325
25 – 34	3,149
35 – 44	1,323
45 – 54	512
55 and over	335
Total	10,392

This kind of frequency distribution is called grouped frequency distribution.

Frequency distributions present data in a relatively compact form, gives 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 abovegrouped frequency distribution cannot tell how many of the arrested persons are 19 years old, or how many are over 62.

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 forma 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) We seldom use fewer than 6 or more than 20 classes; and 15 generally is a good number, the exact number we use in a given situation depends mainly on the number of measurements or observations we have to group

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

$K = 1 + 3.322 \times \log(n)$, where n is the number of observations

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

$W = (\text{Maximum value} - \text{Minimum value})/K = \text{Range}/K$

2) We always make sure that each item (measurement or observation) goes into one and only one class, i.e. classes should be mutually exclusive. To this end we must make sure that the smallest and largest values fall within the classification, that none of the values can fall into possible gaps between successive classes, and that the classes do not overlap, namely, that successive classes have no values in common.

Note that the Sturges rule should not be regarded as final, but should be considered as a guide only. The number of classes specified by the rule should be increased or decreased for convenient or clear presentation.

3) **Determination of class limits**: (i) Class limits should be definite and clearly stated. In other words, open-end classes should be avoided since they make it difficult, or even impossible, to calculate certain further descriptions that may be of interest. These are classes like less then 10, greater than 65, and so on. (ii) The starting point, i.e., the

lower limit of the first class be determined in such a manner that frequency of each class get concentrated near the middle of the class interval. This is necessary because in the interpretation of a frequency table and in subsequent calculation based up on it, the mid-point of each class is taken to represent the value of all items included in the frequency of that class.

It is important to watch whether they are given to the nearest inch or to the nearest tenth of an inch, whether they are given to the nearest ounce or to the nearest hundredth of an ounce, and so forth. For instance, to group the weights of certain animals, we could use the first of the following three classifications if the weights are given to the nearest kilogram, the second if the weights are given to the nearest tenth of a kilogram, and the third if the weights are given to the nearest hundredth of a kilogram:

Weight (kg)	**Weight (kg)**	**Weight (kg)**
10 – 14	10.0 – 14.9	10.00 – 14.99
15 – 19	15.0 – 19.9	15.00 – 19.99
20 – 24	20.0 – 24.9	20.00 – 24.99
25 – 29	25.0 – 29.9	25.00 – 29.99
30 – 34	30.0 – 34.9	30.00 – 34.99

Example: Construct a grouped frequency distribution of the following data on the amount of time (in hours) that 80 college students devoted to leisure activities during a typical school week:

23	24	18	14	20	24	24	26	23	21
16	15	19	20	22	14	13	20	19	27
29	22	38	28	34	32	23	19	21	31
16	28	19	18	12	27	15	21	25	16
30	17	22	29	29	18	25	20	16	11
17	12	15	24	25	21	22	17	18	15
21	20	23	18	17	15	16	26	23	22
11	16	18	20	23	19	17	15	20	10

Using the above formula, K = 1 + 3.322 × log (80) = 7.32 ≈ 7 classes

Maximum value = 38 and Minimum value = 10 ➔ Range = 38 – 10 = 28 and W = 28/7 = 4

Using width of 5, we can construct grouped frequency distribution for the above data as:

Time spent (hours)	Tally	Frequency	Cumulative freq
10 – 14	~~////~~ ///	8	8
15 – 19	~~////~~ ~~////~~ ~~////~~ ~~////~~ ~~////~~ ///	28	36
20 – 24	~~////~~ ~~////~~ ~~////~~ ~~////~~ ~~////~~ //	27	63
25 – 29	~~////~~ ~~////~~ //	12	75
30 – 34	////	4	79
35 – 39	/	1	80

The smallest and largest values that can go into any class are referred to as its class limits; they can be either lower or upper class limits.

For our data of patients, for example
$n = 50$ then $k = 1 + 3.322(\log_{10}50) = 6.64 = 7$ and $w = R / k = (89 - 1)/7 = 12.57 = 13$

Cumulative and Relative Frequencies: When frequencies of two or more classes are added up, such total frequencies are called Cumulative Frequencies. This frequencies help as to find the total number of items whose values are less than or greater than some value. On the other hand, relative frequencies express the frequency of each value or class as a percentage to the total frequency.

Note. In the construction of cumulative frequency distribution, if we start the cumulation from the lowest size of the variable to the highest size, the resulting frequency distribution is called `**Less than cumulative frequency distribution**' and if the cumulation is from the highest to the lowest value the resulting frequency distribution is called `**more than cumulative frequency distribution.'** The most common cumulative frequency is the less than cumulative frequency.

Mid-Point of a class interval and the determination of Class Boundaries

Mid-point or class mark (Xc) of an interval is the value of the interval which lies mid-way between the lower true limit (LTL) and the upper true limit (UTL) of a class. It is calculated as:

$$X_c = \frac{\text{Upper Class Limit} + \text{Lower Class Limit}}{2}$$

True limits (or class boundaries) are those limits, which are determined mathematically to make an interval of a continuous variable continuous in both directions, and no gap exists between classes. The true limits are what the tabulated limits would correspond with if one could measure exactly.

Example: Frequency distribution of weights (in Ounces) of Malignant Tumors Removed from the Abdomen of 57 subjects

Weight	Class boundaries	Xc	Freq.	Cum. freq.	Relative freq (%)
10 - 19	9.5 - 19.5	14.5	5	5	0.0877
20 - 29	19.5 - 29.5	24.5	19	24	0.3333
30 - 39	29.5 - 39.5	34.5	10	34	0.1754
40 - 49	39.5 - 49.5	44.5	13	47	0.2281
50 - 59	49.5 - 59.5	54.5	4	51	0.0702
60 - 69	59.5 - 69.5	64.5	4	55	0.0702
70 - 79	69.5 - 79.5	74.5	2	57	0.0352
Total			57		1.0000

Note: The width of a class is found from the true class limit by subtracting the true lower limit from the upper true limit of any particular class.

For example, the width of the above distribution is (let's take the fourth class) w = 49.5 - 39.5 = 10.

2.7.2 Statistical Tables

A statistical table is an orderly and systematic presentation of numerical data in rows and columns. Rows (stubs) are horizontal and columns (captions) are vertical arrangements. The use of tables for organizing data involves grouping the data into mutually exclusive categories of the variables and counting the number of occurrences (frequency) to each category.

These mutually exclusive categories, for qualitative variables, are naturally occurring groupings. For example, Sex (Male, Female), Marital status (single, Married, divorced, widowed, etc.), Blood group (A, B, AB, O), Method of Delivery (Normal, forceps, Cesarean section, etc.), etc. are some qualitative variables with exclusive categories.

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

Based on the purpose for which the table is designed and the complexity of the relationship, a table could be either of simple frequency table or cross tabulation.

The simple frequency table is used when the individual observations involve only to a single variable whereas the cross tabulation is used to obtain the frequency distribution of one variable by the subset of another variable. In addition to the frequency counts, the relative frequency is used to clearly depict the distributional pattern of data. It shows the percentages of a given frequency count. For simple frequency distributions, (like Table 1) the denominators for the percentages are the sum of all observed frequencies, i.e. 210.

On the other hand, in cross tabulated frequency distributions where there are row and column totals, the decision for the denominator is based on the variable of interest to be compared over the subset of the other variable. For example, in Table 3 the interest is to compare the immunization status of mothers in different marital status group. Hence, the denominators for the computation of proportion of mothers under each marital status group will be the total number of mothers in each marital status category, i.e. row total.

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 orunobserved 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 footnote.

Examples

A) Simple or one-way table: The simple frequency table is used when the individual observations involve only to a single variable whereas the cross tabulation is used to obtain the frequency distribution of one variable by the subset of another variable. In addition to the frequency counts, the relative frequency is used to clearly depict the distributional pattern of data. It shows thepercentages of a given frequency count.

Table 1: Overall immunization status of children in Adami Tullu Woreda, Feb. 1995

Immunization status	Number	Percent
Not immunized	75	35.7
Partially immunized	57	27.1
Fully immunized	78	37.2
Total	210	100.0

Source: Fikru T et al. EPI Coverage in Adami Tulu. Eth J Health Dev 1997;11(2): 109-113

B. Two-way table: This table shows two characteristics and is formed when either the caption or the stub is divided into two or more parts.

In cross tabulated frequency distributions where there are row and column totals, the decision for the denominator is based on the variable of interest to be compared over the subset of the othervariable. For example, in Table 2 the interest is to compare the immunization status of mothers in different marital status group. Hence, the denominators for the computation of proportion of mothers under each marital status group will be the total numbers of mothers in each marital status category, i.e. row total.

Table 2: TT immunization by marital status of the women of childbearing age, Assendabo town, Jimma Zone, 1996

Marital Status	Immunization Status				Total
	Immunized		Non Immunized		
	No.	%	No.	%	
Single	58	24.7	177	75.3	235
Married	156	34.7	294	65.3	450
Divorced	10	35.7	18	64.3	28
Widowed	7	50.0	7	50.0	14
Total	231	31.8	496	68.2	727

Source: Mikael A. et al Tetanus Toxoid immunization coverage among women of child bearing age in Assendabo town; Bulletin of JIHS, 1996, 7(1): 13-20

C. Higher Order Table: When it is desired to represent three or more characteristics in a single table. Thus, if it is desired to represent the `Profession,' `sex' and `Residence,' of the study individuals, the table would take the form as shown in table 3 below and would be called higher order table**.**

Example: A study was carried out on the degree of job satisfaction among doctors and nurses in rural and urban areas. To describe the sample a cross-tabulation was constructed which included the sex and the residence (rural urban) of the doctors and nurses interviewed.

Table 3: Distribution of Health Professional by Sex and Residence

		Residence		
Profession/Sex		Urban	Rural	Total
Doctors	Male	8 (10.0)	35 (21.0)	43 (17.7)
	Female	2 (3.0)	16 (10.0)	18 (7.4)
Nurses	Male	46 (58.0)	36 (22.0)	82 (33.7)
	Female	23 (29.0)	77 (47.0)	100 (41.2)
Total		79 (100.0)	164 (100.0)	243 (100.0)

2.7.3. 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 is specifically interested) would care to study them. Their study is a greater strain upon the mind without, in most cases, any scientific result. The aim of statistical methods, inter alia, is to reduce the size of statistical data and to render them easily intelligible. To attain this objective the methods of classification, tabulation, averages and percentages are generally used. But the method of diagrammatic representation (visual aids) is probably simpler and more easily understandable. It consists in presenting statistical material in geometric figures, pictures, maps and lines or curves.

Importance of Diagrammatic Representation

1. They have greater attraction than mere figures. They give delight to the eye and add a spark of interest.
2. They help in deriving the required information in less time and without any mental strain.
3. They facilitate comparison.
4. They may reveal unsuspected patterns in a complex set of data and may suggest directions in which changes are occurring. This warns us to take an immediate action.
5. They have greater memorising value than mere figures. This is so because the impression left by the diagram is of a lasting nature.

Limitations of Diagrammatic Representation

1. The technique of diagrammatic representation is made use only for purposes of comparison. It is not to be used when comparison is either not possible or is not necessary.

2. Diagrammatic representation is not an alternative to tabulation. It only strengthens the textual exposition of a subject, and cannot serve as a complete substitute for statistical data.
3. It can give only an approximate idea and as such where greater accuracy is needed diagrams will not be suitable.
4. They fail to bring to light small differences

Construction of graphs

The choice of the particular form among the different possibilities will depend on personal choices and/or the type of the data.

- Bar charts and pie chart are commonly used for qualitative or quantitative discrete data.
- Histograms, frequency polygons are used for quantitative continuous data.

There are, however, general rules that are commonly accepted about construction of graphs.

1. Every graph should be self-explanatory and as simple as possible.
2. Titles are usually placed below the graph and it should again question what ? Where? When? How classified?
3. Legends or keys should be used to differentiate variables if more than one is shown.

4. The axes label should be placed to read from the left side and from the bottom.
5. The units in to which the scale is divided should be clearly indicated.
6. The numerical scale representing frequency must start at zero or a break in the line should be shown.

Examples of diagrams:

1. Bar Chart

Bar diagrams are used to represent and compare the frequency distribution of discrete variables and attributes or categorical series. When we represent data using bar diagram, all the bars must have equal width and the distance between bars must be equal.

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.

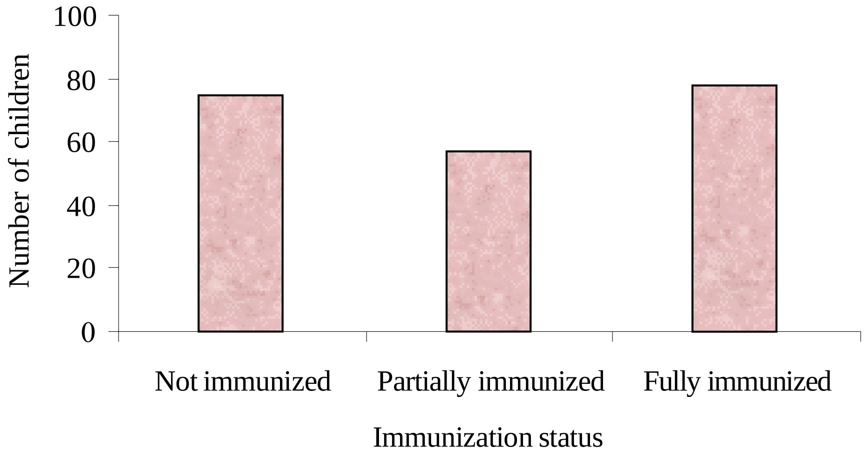


Fig. 1. Immunization status of Children in Adami Tulu Woreda, Feb. 1995

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 of multiple bar diagrams: consider that data on immunization status of women by marital status.

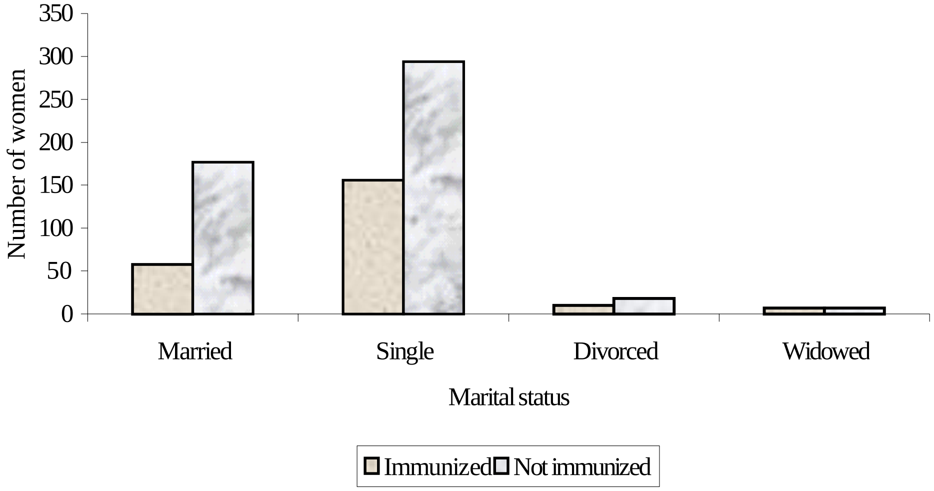


Fig. 2 TT Immunization status by marital status of women 15-49 years, Asendabo town, 1996

C. Component (or sub-divided) Bar Diagram: Bars are sub-divided into component parts of the figure. These sorts of diagrams are constructed when each total is built up from two or more component figures. They can be of two kind:

I) Actual Component Bar Diagrams: When the over all height of the bars and the individual component lengths represent actual figures.

Example of actual component bar diagram: The above data can also be presented as below.

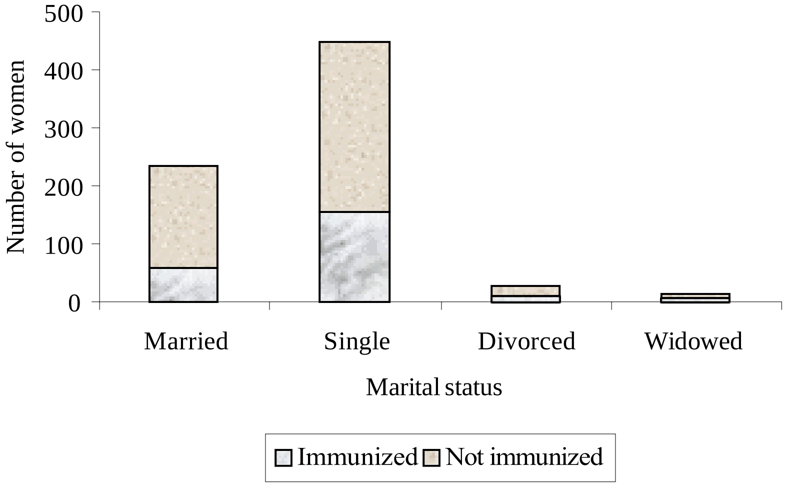


Fig. 3 TT Immunization status by marital status of women 15-49 years, Asendabo town, 1996

ii) Percentage Component Bar Diagram: Where the individual component lengths represent the percentage each component forms the over all total. Note that a series of such bars will all be the same total height, i.e., 100 percent.

Example of percentage component bar diagram:

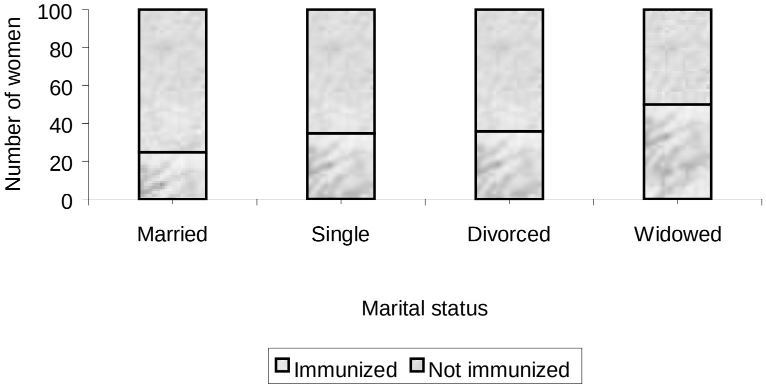


Fig. 4 TT Immunization status by marital status of women 15-49years, Asendabo town, 1996

2) Pie-chart (qualitative or quantitative discrete data): it is a circle divided into sectors so that the areas of the sectors are proportional to the frequencies.

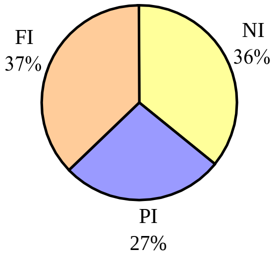


Fig. 5. Immunization status of children in Adami Tullu Woreda, Feb. 1995

3. Histograms (quantitative continuous data)

A histogram is the graph of the frequency distribution of continuous measurement variables. It is constructed on the basis of the following principles:

a) The horizontal axis is a continuous scale running from one extreme end of the distribution to the other. It should be labelled with the name of the variable and the units of measurement.

b) For each class in the distribution a vertical rectangle is drawn with (i) its base on the horizontal axis extending from one class boundary of the class to the other class boundary, there will never be any gap between the histogram rectangles. (ii) the bases of all rectangles will be determined by the width of the class intervals. If a distribution with unequal class-interval is to be presented by means of a histogram, it is necessary to make adjustment for varying magnitudes of the class intervals.

Example: Consider the data on time (in hours) that 80 college students devoted to leisure activities during a typical school week:

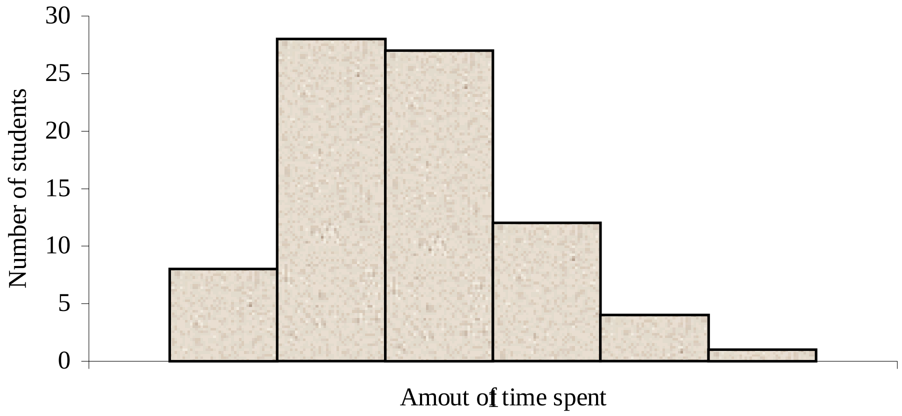


Fig 6: Histogram for amount of time college students devoted to leisure activities

2. FREQUENCY POLYGON:

If we join the midpoints of the tops of the adjacent rectangles of the histogram with line segments a frequency polygon is obtained. When the polygon is continued to the X-axis just out side the range of the lengths the total area under the polygon will be equal to the total area under the histogram.

Note that it is not essential to draw histogram in order to obtain frequency polygon. It can be drawn with out erecting rectangles of histogram as follows:

1) The scale should be marked in the numerical values of the mid-points of intervals.
2) Erect ordinates on the midpoints of the interval - the length or altitude of an ordinate representing the frequency of the class on whose mid-point it is erected.
3) Join the tops of the ordinates and extend the connecting lines to the scale of sizes.

Example: Consider the above data on time spend on leisure activities

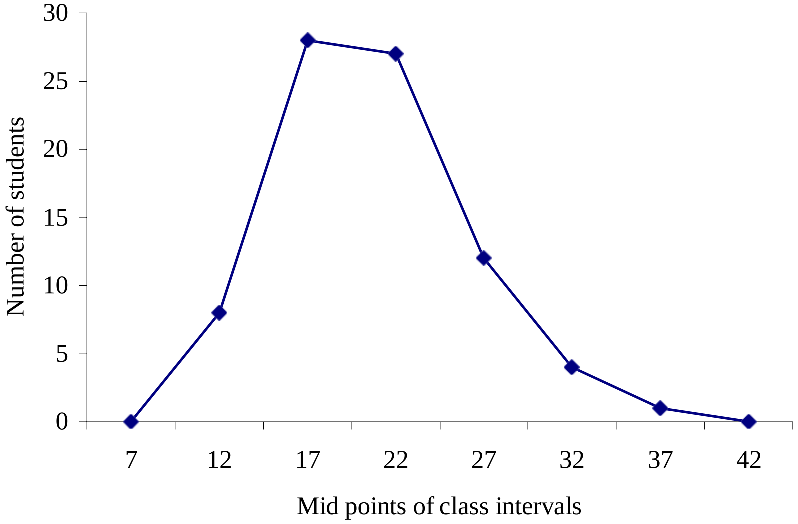


Fig 7: Frequency polygon curve on time spent for leisure activities by students

3. **OGIVE OR CUMULATIVE FREQUENCY CURVE:** When the **cumulative frequencies** of a distribution are graphed the resulting curve is called **Ogive Curve**.

To construct an Ogive curve:

i) Compute the cumulative frequency of the distribution.

ii) Prepare a graph with the cumulative frequency on the vertical axis and the true upper class limits (class boundaries) of the interval scaled along the X-axis (horizontal axis). The true lower limit of the lowest class interval with lowest scores is included in the X-axis scale; this is also the true upper limit of the next lower interval having a cumulative frequency of 0.

Example: Consider the above data on time spend on leisure activities

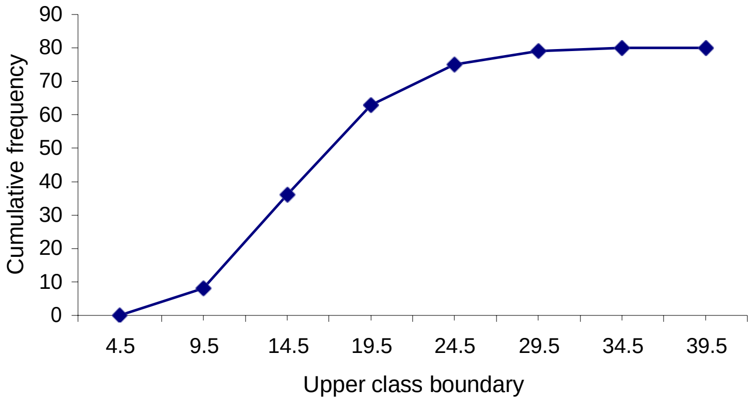


Fig 8: Cumulative frequency curve for amount of time college students devoted to leisure activities

4. The line diagram

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. The line graph is suitable for depicting a consecutive trend of a series over a long period.

Example: Malaria parasite rates as obtained from malaria seasonal blood survey results, Ethiopia (1967-79 E.C)

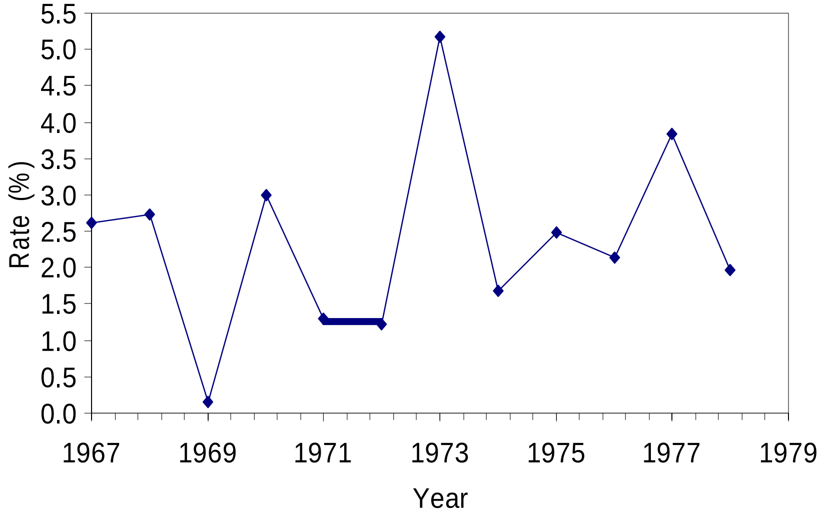


Fig 9: Malaria Parasite Prevalence Rates in Ethiopia, 1967 – 1979 Eth. C.

CHAPTER THREE

Summarizing Data

3.1. Learning objectives

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

1. Identify the different methods of data summarization
2. Compute appropriate summary values for a set of data
3. Appreciate the properties and limitations of summary values

3.2. 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.

The basic problem of statistics can be stated as follows: Consider a sample of data X_1, X_n, where X_1 corresponds to the first sample point and X_n corresponds to the nth sample point. Presuming that the sample is drawn from some population P, what inferences or conclusion can be made about P from the sample?

Before this question can be answered, the data must be summarized as succinctly (concisely, briefly) as possible, since the number of sample points is frequently large and it is easy to lose track of the overall picture by looking at all the data at once. One type of measure useful for summarizing data defines the center, or middle, of the sample. This type of measure is a measure of central tendency (location).

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

Notations: $\sum$ is read as Sigma (the Greek Capital letter for S) means the sum of

Suppose n values of a variable are denoted as $x_1, x_2, x_{3\ldots,} x_n$ then $\sum x_i = x_1, + x_2, + x_3 + \ldots x_n$ where the subscript i range from 1 up to n

Example: Let $x_1=2$, $x_2 = 5$, $x_3=1$, $x_4 = 4$, $x_5=10$, $x_6= -5$, $x_7 = 8$

Since there are 7 observations, i range from 1 up to 7

i) $\sum x_i = 2+5+1+4+10-5+8 = 25$

ii) $(\sum x_I)^2 = (25)^2 = 625$

iii) $\sum x_I^2 = 4 + 25 + 1 + 16 + 100 + 25 + 64 = 235$

Rules for working with summation

1) $\sum(x_i + y_i) = \sum x_i + \sum y_i$, where the number of x values = the number of y values.
2) $\sum Kx_i = k \times \sum x_i$, where K is a constant.
3) $\sum K = n \times K$, where K is a constant.

3.3. Measures of Central Tendency

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.

1. The Arithmetic Mean or simple Mean

Suppose the sample consists of birth weights (in grams) of all live born infants born at a private hospital in a city, during a 1-week period. This sample is shown in the following table:

3265	3323	2581	2759	3260	3649	2841
3248	3245	3200	3609	3314	3484	3031
2838	3101	4146	2069	3541	2834	

One measure of central location for this sample is the arithmetic mean ; it is usually denoted by $\overline{X}$.

Definition: the arithmetic mean is the sum of all observations divided by the number of observations. It is written in statistical terms as:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{ , i = 1, 2, …, n}$$

Example: What is the arithmetic mean for the sample birth weights?

$$\overline{X} = \frac{1}{n}\Sigma x_i = \frac{1}{20} \text{ (3265 + 3260 + ….+ 2834) = } \frac{63{,}338}{20} = 3166.9 \text{ g}$$

The arithmetic mean is, in general, a very natural measure of central location. One of its principal limitations, however, is that it is **overly sensitive to extreme values**. In this instance it may not be representative of the location of the great majority of the sample points.

For example, if the first infant were in the above data happened to be a premature infant weighing 500 gm rather than 3265 g, then the arithmetic mean of the sample would be reduced to 3028.7 g. In this instance, 7 of the birth weights would be lower the arithmetic mean, and 13 would be higher than the mean. It is possible in extreme cases for all but one of the sample points to be on one side of the arithmetic mean. The arithmetic mean is a poor measure of central location in these types of sample, since it does not reflect the center of sample. Nevertheless, the arithmetic mean is by far the most widely used measure of central location.

2. Median

An alternative measure of central location, perhaps second in popularity to the arithmetic mean, is the median.

Suppose there are n observations in a sample. If these observations are ordered from smallest to largest, then the median is defined as follows:

Definition: The sample median is

(1) The $\left(\frac{n+1}{2}\right)^{th}$ observations if n is odd

(2) The average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations if n is even.

The rational for these definitions is to ensure an equal number of sample points on both sides of the sample median.

The median is defined differently when n is even and odd because it is impossible to achieve this goal with one uniform definition. For samples with an add sample size, there is a unique central point; for example, for sample of size 7, the fourth largest point is the central point in the sense that 3 points are both smaller and larger than it. For samples with an even size, there is no unique central point and the middle 2 values must be averaged. Thus, for sample of size 8, the fourth and the fifth largest points would be averaged to obtain the median, since neither is the central point.

Example: Compute the sample median for the birth weight data

Solution: First arrange the sample in ascending order
2069, 2581, 2759, 2834, 2838, 2841, 3031, 3101, 3200, 3245, 3248, 3260, 3265, 3314, 3323, 3484, 3541, 3609, 3649, 4146

Since n=20 is even,
Median = average of the 10th and 11th largest observation = (3245 + 3248)/2 = 3246.5 g

Example: Consider the following data, which consists of white blood counts taken on admission of all patients entering a small hospital on a given day. Compute the median white-blood count ($\times 10^3$). 7, 35,5,9,8,3,10,12,8

Solution: First, order the sample as follows. 3,5,7,8,8,9,10,12,35. Since n is odd, the sample median is given by the 5th, ((9+1)/2)th, largest point, which is equal to 8.
The principal strength of the sample median is that it is **insensitive to very large or very small values**.

In particular, if the second patient in the above data had a white blood count of 65,000 rather than 35,000, the sample median would remain unchanged, since the fifth largest value is still 8,000. Conversely the arithmetic mean would increase dramatically from 10,778 in the original sample to 14,111 in the new sample.

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

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. The mode is not often used in biological or medical data.

Find the modal values for the following data

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

b) 1.8, 3.0, 3.3, 2.8, 2.9, 3.6, 3.0, 1.9, 3.2, 3.5 (modal value = 3.0 kg)

Skewness: If extremely low or extremely high observations arepresent 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.

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.

c) **Symmetrical 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.

In unimodal (one-peak) symmetrical distributions, the mean, median and mode are identical. On the other hand, in unimodal skewed distributions, it is important to remember that the mean, median and mode occur in alphabetical order when the longer tail is at the left of the distribution or in reverse alphabetical order when the longer tail is at the right of the distribution.

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

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 nth 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.

Examle: 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

$$GM = \sqrt[42]{7x8x3x\ldots x1x237}$$

$$\log Gm = 1/42\ (\log 7+\log 8+\log 3+..+\log 237)$$

$$= 1/42\ (.8451+.9031+.4771+\ldots 2.3747)$$

$$= 1/42\ (41.9985)$$

$$= 0.9999 \approx 1.0000$$

The anti-log of 0.9999 is 9.9992 $\approx$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 $\approx$ 40.

Summary

The process of computing any one of the averages discussed so far is comparatively simple. But it is not always easy to choose one particular average which may represent a statistical distribution for the purpose of the inquiry that we have in hand. Below is given a summary of the characteristics, advantages and disadvantages of each average in order to enlarge the awareness of the user so that the selection process could be facilitated.

A) Mean

i) Characteristics:-

1) The value of the arithmetic mean is determined by every item in the series.
2) It is greatly affected by extreme values.
3) The sum of the deviations about it is zero.
4) The sum of the squares of deviations from the arithmetic mean is less than of those computed from any other point.

ii) Advantages

1) It is based on all values given in the distribution.
2) It is most early understood.
3) It is most amenable to algebraic treatment.

iv) Disadvantages

1) It may be greatly affected by extreme items and its usefulness as a “Summary of the whole” may be considerably reduced.
2) When the distribution has open-end classes, its computation would be based assumption, and therefore may not be valid.

B) Median

I

i) Characteristics

1) It is an average of position.
2) It is affected by the number of items than by extreme values.

ii) Advantages

1) It is easily calculated and is not much disturbed by extreme values
2) It is more typical of the series
3) 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.

iii) Disadvantages

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

C. Mode

i) Characteristics

1) It is an average of position
2) It is not affected by extreme values
3) It is the most typical value of the distribution

iii) Advantages

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

iii) Disadvantages:-

1. Unless the number of items is fairly large and the distribution reveals a distinct central tendency, the mode has no significance
2. It is not capable of mathematical treatment
3. In a small number of items the mode may not exist.

D. Geometric Mean

i) Characteristics

1. It is a calculated value and depends upon the size of all the items.
2. It gives less importance to extreme items than does the arithmetic mean.

3. For any series of items it is always smaller than the arithmetic mean.
4. It exists ordinarily only for positive values.

ii) Advantages:-

1) since it is less affected by extremes it is a more preferable average than the arithmetic mean
2) It is capable of algebraic treatment
3) It based on all values given in the distribution.

iv) Disadvantages:-

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

3.4. 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 abut the center.

Consider the following data sets:

								Mean
Set 1:	60	40	30	50	60	40	70	50
Set 2:	50	49	49	51	48	50	53	50

The two data sets given above have a mean of 50, but obviously set 1 is more “spread out” than set 2. How do we express this numerically? The object of measuring this scatter or dispersion is to obtain a single summary figure which adequately exhibits whether the distribution is compact or spread out.

Some of the commonly used measures of dispersion (variation) are: Range, interquartile range, variance, standard deviation and coefficient of variation.

1. Range

The range is defined as the difference between the highest and smallest observation in the data. It is the crudest measure of dispersion. The range is a measure of absolute dispersion and as such cannot be usefully employed for comparing the variability of two distributions expressed in different units.

$$\text{Range} = x_{max} - x_{min}$$

Where , $x_{max} =$ highest (maximum) value in the given distribution.

Xmin = lowest (minimum) value in the given distribution.

In our example given above (the two data sets)

* The range of data in set 1 is 70-30 =40

* The range of data in set 2 is 53-48 =5

1. Since it is based upon two extreme cases in the entire distribution, the range may be considerably changed if either of the extreme cases happens to drop out, while the removal of any other case would not affect it at all.

2. It wastes information for it takes no account of the entire data.

3. The extremes values may be unreliable; that is, they are the most likely to be faulty

4. Not suitable with regard to the mathematical treatment required in driving the techniques of statistical inference.

2. Quantiles

Another approach that addresses some of the shortcomings of the range is in quantifying the spread in the data set is the use of quantiles or percentiles. Intuitively, the pth percentile is the value Vp such that p percent of the sample points are less than or equal to Vp. The median, being the 50th percentile, is a special case of a quantile. As was the case for the median, a different definition is needed for the pth percentile, depending on whether np/100 is an integer or not.

Definition: The pth percentile is defined by

(1) The $(k+1)^{th}$ largest sample point if np/100 is not an integer (where k is the largest integer less than np/100)

(2) The average of the $(np/100)^{th}$ and $(np/100 + 1)^{th}$ largest observation is np/100 is an integer.

The spread of a distribution can be characterized by specifying several percentiles. For example, the 10th and 90th percentiles are often used to characterize spread. Percentages have the advantage over the range of being less sensitive to outliers and of not being much affected by the sample size (n).

Example: Compute the 10th and 90th percentile for the birth weight data.

Solution: Since $20\times0.1=2$ and $20\times0.9=18$ are integers, the 10th and 90th percentiles are defined by

10th percentile = the average of the 2nd and 3rd largest values = (2581+2759)/2 = 2670 g

90th percentile=the average of the18th and 19th largest values = (3609+3649)/2 = 3629 grams.

We would estimate that 80 percent of birth weights would fall between 2670 g and 3629 g, which gives us an overall feel for the spread of the distribution.

Other quantlies which are particularly useful are the **quartiles** of the distribution. The quartiles divide the distribution into four equal parts. The second quartile is the median. The interquartile range is the difference between the first and the third quartiles.

To compute it, we first sort the data, in ascending order, then find the data values corresponding to the first quarter of the numbers (first quartile), and then the third quartile. The interquartile range (IQR) is the distance (difference) between these quartiles.

Eg. Given the following data set (age of patients):-

18,59,24,42,21,23,24,32

find the interquartile range!

1. sort the data from lowest to highest
2. find the bottom and the top quarters of the data
3. find the difference (interquartile range) between the two quartiles.

18 21 23 24 24 32 42 59

1st quartile = The {(n+1)/4}th observation = (2.25)th observation

= 21 + (23-21)x .25 = 21.5

3rd quartile = {3/4 (n+1)}th observation = (6.75)th observation

= 32 + (42-32)x .75 = 39.5

Hence, IQR = 39.5 - 21.5 = 18

The interquartile range is a preferable measure to the range. Because it is less prone to distortion by a single large or small value. That is, outliers in the data do not affect the inerquartile range. Also, it can be computed when the distribution has open-end classes.

3. Standard Deviation and Variance

Definition: The sample and population standard deviations denoted by S and σ (by convention) respectively are defined as follows:

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

$$\sigma = \sqrt{\frac{\sum (X_i - \mu)^2}{N}}$$ = population standard deviation

This measure of variation is universally used to show the scatter of the individual measurements around the mean of all the measurements in a given distribution.

Note that the sum of the deviations of the individual observations of a sample about the sample mean is always 0.

The square of the standard deviation is called the variance. The variance is a very useful measure of variability because it uses the information provided by every observation in the sample and also it is very easy to handle mathematically. Its main disadvantage is that the units of variance are the square of the units of the original observations. Thus if the original observations were, for example, heights in cm then the units of variance of the heights are cm^2. The easiest way around this difficulty is to use the square root of the variance (i.e., standard deviation) as a measure of variability.

Computational formulas for the sample variance or SD

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - n\mathrm{X}^2}{n-1} \quad \text{and } S^2 = \frac{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}{n-1}$$

S = $\sqrt{\dfrac{\sum\limits_{i=1}^{n} X_i - n\overline{X}^2}{n-1}}$ and S = $\sqrt{\dfrac{n\sum\limits_{i=1}^{n} X_i^2 - (\sum\limits_{i=1}^{n} X_i)^2}{n(n-1)}}$

Example: Areas of sprayable surfaces with DDT from a sample of 15 houses are as follows (m^2) :

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

Find the variance and standard deviation of the above distribution.

The mean of the sample is 125 m^2.

Variance (sample) = $s^2 = \sum(xi - x)^2/n-1$

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

= 2502/14

= 178.71 (square metres)^2

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

Some important properties of the arithmetic mean and standard deviation

Consider a sample X_1, ….., X_n, which will be referred to as the original sample. To create a translated sample X_1+c, add a constant C to each data point. Let $y_i = x_i + c$, i = 1, …., n. Suppose we want to compute the arithmetic mean of the translated sample, we can show that the following relationship holds:

1. If $Y_i = X_i + c$, i = 1, …., n

Then $\overline{Y} = \overline{X}$ + c and Sy = Sx

Therefore, to find the arithmetic mean of the Y's, compute the arithmetic mean of the X's and add the constant c but the standard deviation of Y will be the same as the standard deviation of X.

This principle is useful because it is sometimes convenient to change the "origin" of the sample data, that is, compute the arithmetic mean after the translation and transform back to the original data.

What happens to the arithmetic mean if the units or scales being worked with are changed? A re-scaled sample can be created:

2. If $Y_i = cx_i$, i=1, ……., n

Then $\overline{Y} = c\overline{X}$ and Sy = cSx

Therefore, to find the arithmetic mean and standard deviation of the Y's compute the arithmetic mean and standard deviation of the X's and multiply it by the constant c.

Example: Express the mean and standard deviation of birth weight for the above data in ounces rather than grams.

We know that 1 oz = 28.35 gm and that $\overline{X}$ = 3166.9 gm and S_x = 445.3 gm. Thus, if the data were expressed in terms of ounces: $c = \frac{1}{28.35}$ and $\overline{Y} = \frac{1}{28.35}$(3166.9) = 111.71 oz and

S_y = 1/28.35(445.3) = 15.7 oz

Sometimes we want to change both the origin and the scale of the data at the same time. To do this,

Let X_1, ….. Xn be the original sample of the data and let $Y_i = c_1X_i + c_2$,i = 1, …, n, represent a transformed sample obtained by multiplying each original sample point by a factor c_1, and then shifting over by a constant c_2

3. If $Y_i = c_1x_i + c_2$, i = 1, ……, n

Then $\overline{Y} = c_1\overline{X} + c_2$ and $S_y = c_1S_x$

If we have a sample of temperature in °C with an arithmetic mean of 11.75° with standard deviation of 1.8, then what is the arithmetic mean and standard deviation in °F?

Let Y_i denote the °F temperature that corresponds to a °Ctemperature of X_i. Since the required transformation to convert the data to °F would be

$Y_i = \frac{9}{5}X_i + 32$, i = 1, ….., n

Then the arithmetic mean and standard deviation in ^{0}F would be

$$\bar{Y} = \frac{9}{5}(11.75) + 32 = 53.15°F \text{ and } Sy = 9/5(1.8) = 3.24\ ^{0}F$$

Weighted Mean of Sample Means and Pooled Standard Deviation

When averaging quantities, it is often necessary to account for the fact that not all of them are equally important in the phenomenon being described. In order to give quantities being averaged there proper degree of importance, it is necessary to assign them relative importance called ***weights***, and then calculate a weighted mean. In general, the weighted mean $\bar{X}_w$ of a set of numbers X_1, X_2, … and X_n, whose relative importance is expressed numerically by a corresponding set of numbers w_1, w_2, … and w_n, is given by

$$\bar{X}_w = \frac{w_1X_1 + w_2X_2 + \ldots + w_nX_n}{w_1 + w_2 + \ldots + w_n} = \frac{\sum w \times X}{\sum w}$$

Example: In a given drug shop four different drugs were sold for unit price of 0.60, 0.85, 0.95 and 0.50 birr and the total number of drugs sold were 10, 10, 5 and 20 respectively. What is the average price of the four drugs in this drug shop?

Solution: for this example we have to use weighted mean using number of drugs sold as the respective weights for each drug's price. Therefore, the average price will be:

$$\overline{X}_w = \frac{10\times 0.60 + 10\times 0.85 + 5\times 0.95 + 20\times 0.50}{10+10+5+20} = \frac{29.25}{45} = 0.65 \text{ birr}$$

If we don't consider the weights, the average price will be 0.725 birr

On the other hand, where several means ($\overline{X}$'s) and standard deviations (s's) for a variable are available and if we need to compute the over all mean and standard deviation, the weighted mean ($\overline{X}_w$) and pooled standard deviation (Sp) of the entire group consisting of all the samples may be computed as:

$$\overline{X}_w = \frac{n_1\overline{X}_1 + n_2\overline{X}_2 + \ldots + n_k\overline{X}_k}{n_1 + n_2 + \ldots + n_k}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2 + \ldots + (n_k-1)S_k^2}{(n_1-1)+(n_2-1)+\ldots+(n_k-1)}}$$

where n_i , $\overline{X}_i$ and S_i represent number of observations, mean and standard deviation of each single sample, respectively.

Example: The mean systolic blood pressure was found to be 129.4 and 133.6 mm Hg with standard deviations of 10.6 and 15.2 mm Hg, respectively, for two groups of 12 and 15 men. What is the mean systolic pressure of all the 27 men?

Solution: Given: Group 1: $\overline{X}_1 = 129.4$, $S_1 = 10.6$ and $n_1 = 12$

Group 2: $\overline{X}_2 = 133.6$, $S_2 = 15.2$ and $n_2 = 15$

The mean of the 27 men is given by the weighted mean of the two groups.

$$\overline{X}_W = \frac{n_1\overline{X}_1 + n_2\overline{X}_2}{n_1 + n_2} = \frac{12(129.4) + 15(133.6)}{12 + 15} = 131.73 \text{ mm Hg}$$

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}} = \sqrt{\frac{11\times 10.6^2 + 14\times 15.2^2}{11+14}} = 13.37 \text{ mm Hg}$$

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 and SD together, since, for example, a standard deviation of 10 would mean somethingdifferent 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 (CV) is defined by: $100\% \times \frac{S}{\overline{X}}$

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, because if the units are changed by a factor C, both the mean and SD change by the factor C; the CV, which is the ratio between them, remains uncharged.

Example: 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

$$CV = 100\% \times \frac{S}{\overline{X}} = 100\% \times \frac{15.7}{111.71} = 14.1\%$$

Computation of Summary values from Grouped Frequency Distribution:

Example: Consider the following grouped data on the amount of time (in hours) that 80 college students devoted to leisure activities during a typical school week:

TIME (HOURS)	NO. OF STUDENTS (1)	CUM. FREQ.	MID-POINT (2)	(1)×(2)	(1)×(2)^2	d	f×d	f× d^2
10–14	8	8	12	96	1,152	-2	-16	32
15–19	28	36	17	476	8,092	-1	-28	28
20–24	27	63	22	594	13,068	0	0	0
25–29	12	75	27	324	8,748	1	12	12
30–34	4	79	32	128	4,096	2	8	16
35–39	1	80	37	37	1,369	3	3	9
Total	80			1,655	36,525		-21	97

a) Computing the arithmetic mean

i) **By Direct Method**: This method is applicable where the entire range of available values or scores of the variable has been divided into equal or unequal class intervals and the observations have been grouped into a frequency distribution on that basis. The value or score (x) of each observation is assumed to be identical with the mid-point (X_c) of the class interval to which it belongs. In such cases the mean of the distribution is computed as:

$$\overline{X} = \frac{\sum fX_c}{\sum f} = \frac{\sum fX_c}{n}$$

For the time data the mean time spent by students for leisure activities was:

$$\overline{X} = \frac{\sum fX_c}{\sum f} = \frac{1,655}{80} = 20.7 hours$$

ii) **Indirect or Code Method:** This is applicable only for grouped frequency distribution with equal class interval length.

Steps:

1. Assume an arbitrary mid-point (A) as the mean of the distribution and give a code or coded value of 0.
2. Code numbers (d),..., -2, -1 and 1, 2,... are then assigned in descending and ascending order to the mid-points of the intervals running downwards from and rising progressively higher than A, respectively. The code numbers (d) of the mid-point (Xc) of a class interval may also be obtained by the following way. d = (Xc-A)/w.
3. The code numbers (d) of each class interval is multiplied by the frequency (f) of that interval and the sum of these products (Σfd) is divided by the total number of observations (n) of the sample to get the mean of the coded values;

$\overline{X} = A + \overline{d}w = A + \frac{\sum fd}{n} \times w$ (Remember properties of mean)

From the above data, let A = 22 mid-point of the 3rd class

Σfd = -21, ➔ $\overline{d} = \frac{-21}{80} = 0.2625$

$$\bar{X} = A + \frac{\Sigma fd}{n} \times w = 22 + \frac{-21}{80} \times 5 = 22 + 1.3125 = 20.7 hours$$

b) Computation of Median from a Grouped Frequency Distribution

For a continuous frequency distribution, median is calculated as

$$\tilde{X} = l + \frac{(\frac{1}{2}n - F_l)}{f_{50}} \times w$$

Where l = true lower limit of the interval containing the median, i.e., the median class

w = length of the interval,

n = total frequency of the sample

F_l = Cumulative frequency of all interval below l.

½n = Number of observations to be counted off from one end of the distribution to reach the median and

f_{50} = Frequency of that interval containing the median.

The median class is first class whose cumulative frequency is at least $\left(\frac{n}{2}\right)$.

For the above data, the median class = the first class with $\left(\frac{n}{2}\right)^{th}$ cumulative frequency = the first class whose cumulative frequency is at least 40.

The class whose cumulative frequency at least 40 is the 3rd class, i.e. 20 - 24, for this class then:

LCB = 19.5, frequency of the median class = 27, cumulative frequency next below the median class =36

$$\tilde{X} = 19.5 + \left(\frac{40 - 36}{27} \right) \times 5 = 19.5 + 0.7 = 20.2 \text{ hours}$$

In the calculation of the median from a grouped frequency table, the basic assumption is that within each class of the frequencydistribution, observations are uniformly or evenly distributed over the class interval.

c) Computation of the standard deviation from a Grouped Frequency Distribution

In the calculation of the median from a frequency table, the basic assumption is that within each class of the frequency distribution, observations are uniformly or evenly distributed over the class interval. The frequencies are arranged in a cumulative frequency distribution to facilitate computations.

In a grouped frequency distribution, the SD is computed as

$$S == \sqrt{\frac{80(36{,}525) - (1{,}655)^2}{80(80-1)}} = 5.38 \text{ hours}$$

Using the code method, the SD for equal class interval grouped frequency distribution can be calculated as

$$S = W\times\sqrt{\frac{\Sigma f(d-\bar{d})^2}{n-1}} = W\times\sqrt{\frac{n\Sigma fd^2-(\Sigma fd)^2}{n(n-1)}}$$

Remember characteristics of standard deviation that SD of a constant is 0.

For the above data, $\Sigma fd = -21$, $\Sigma fd^2 = 97$, $W = 5$

$$S = 5 \times\sqrt{\frac{80(97)-(-21)^2}{80(80-1)}} = 5 \times 1.076 = 5.38 \text{ hours}$$

$$\sqrt{\frac{\Sigma f_i(X_{ci}-\overline{X})^2}{\Sigma f_i -1}} = \sqrt{\frac{\Sigma f_i(X_{ci}-\overline{X})^2}{n-1}} = \sqrt{\frac{n\Sigma f_i X_{ci}^2-(\Sigma f_i X_{ci})^2}{n(n-1)}}$$

Where Xci is the mid-point of the ith class.

Example: Consider the previous data on time spend by college students for leisure activities

$$S = \sqrt{\frac{80(36{,}525)-(1{,}655)^2}{80(80-1)}} = 5.38 \text{ hours}$$

Using the code method, the SD for equal class interval grouped frequency distribution can be calculated as

S =W× $\sqrt{\frac{\Sigma f(d-\overline{d})^2}{n-1}}$ = W× $\sqrt{\frac{n\Sigma fd^2-(\Sigma fd)^2}{n(n-1)}}$

Remember characteristics of standard deviation that SD of a constant is 0.

For the above data, $\Sigma fd = -21$, $\Sigma fd^2 = 97$, W = 5

S = 5 × $\sqrt{\frac{80(97)-(-21)^2}{80(80-1)}}$ = 5 × 1.076 = 5.38 hours

3.5 Exercises

Find the mean, median and standard deviation of the following grouped data.

1. Total circulating albumin in gm for 30 normal males, age 20-29 years

Circulating albumin in gm (CI)	Frequency (f)
100-109	2
110-119	6
120-129	6
130-139	7
140-149	8
150-159	1
Total	30

2. Blood pressure levels of 60 first-year male medical students (in mm Hg)

Class limit	Frequency
90-99	2
100-109	6
110-119	17
120-129	16
130-139	12
140-149	6
150-159	1

CHAPTER FOUR

DEMOGRAPHY AND HEALTH SERVICES STATISTICS

4.1. LEARNING OBJECTIVES

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

1. Define and understand the concepts of demographic statistics
2. Identify the different methods of data collection for demographic studies
3. Understand different ratios used to describe demographic data
4. Understand fertility and mortality measures
5. Understand methods of population projection and computation of doubling time
6. Understand and compute the different indices relating to hospitals

4.2. INTRODUCTION

Definition: Demography is a science that studies human population with respect to size, distribution, composition, social mobility and its variation with respect to all the above features and the causes of such variation and the effect of all these on health, social, ethical, and economic conditions.

Size: is the number of persons in the population at a given time.
Example: The size of Ethiopian population in 2002 is about 65 million.

Distribution: is the arrangement of the population in the territory of the nation in geographical, residential area, climatic zone, etc.
Example: Distribution of Ethiopian population by Zone

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**.

Therefore, demographic statistics is the application of statistics to the study of human population in relation to the essential demographic variables and the source of variations of the population with respect to these variables, such as fertility, mortality and migration. It is indispensable to study these variables through other socio-economic variables such as place of birth, religion, occupation, education, marital status, etc.

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 periodof 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.

I) THE CENSUS

In modern usage, the term “census” refers to a nation-wide counting of population. It is obtained by a direct canvass of each person or household, which 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.

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:

a) advantages

- offers less chance of double counting
- Gives less chance for the omission of persons from the count

b) disadvantages:

- Population figures may be inflated or deflated by tourists, travelling salesmen, and other transients.
- In areas with high migration, the registration of vital events (vital statistics) is subject (liable) to distortion.

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 characteristics, 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 summarized as follows:

Individual enumeration, universality within a defined territory, simultaneity and defined periodicity

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

1) gives complete and valid picture of the population composition and characteristics
2) serves as a sampling frame
3) provides with vital statistics of the population in terms of fertility and mortality.
4) Census data are utilized in a number of ways for planning the welfare of the people

Eg. To ascertain food requirements, to plan social welfare schemes like schools, hospitals, houses, orphanages, pensions, etc.

Common errors in census data

1. Omission and over enumeration.

2. Misreporting of age due to memory lapse, preference of terminal digits, over/under estimation.
3. Overstating of the status within the occupation.
4. Under reporting of births due to problem of reference period and memory lapse.
5. Under reporting of deaths due to memory lapse and tendency not to report on deaths, particularly on infant deaths.

ii) Surveys: A survey is a technique based on sampling methods by means of which we try to obtain specific information from part of the population liable to be considered as representative of the whole. Surveys are made at a given moment, in a specific territory; sporadically and without periodicity for the deep study of a problem.

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

Changes in population numbers are taking place every day. Additions are made by births or through new arrivals from outside the area. Reductions take place because of deaths, or through people leaving the area.

The taking of a census is merely a snapshot of an event while the counting of births and deaths (vital records) is a continuous process. The registration of vital events (births, deaths, marriages, etc) is a system by which all births, deaths, etc. occurring nationwide are registered, reported to a control body and compiled centrally. A certificate is issued for every death and birth. The four main characteristics of vital registration are comprehensiveness, compulsory by law, compilation done centrally and the registration is an ongoing (continuous) process.

This system is not well developed in Ethiopia and hence cannot serve as a reliable means of getting information.

4.4 Stages in Demographic Transition

Demographic transition is a term used to describe the major demographic trends of the past two centuries. The change in population basically consists of a shift from an equilibrium condition of high birth and death rates, characteristics of agrarian societies, to a newer equilibrium in which both birth and death rates are at much lower levels. The end stage (period 3) of this demographic transition is a situation in which birth and death rates are again essentially in balance but a lower level.

The theory of demographic transition attempts to explain the changes in mortality and fertility in three different stages.

1) pre-transitional :- characterized by high mortality and high fertility, with low (moderate) population growth (young population). This is shown by the triangular, broad based pattern reflecting the high birth rates, over a long period of time. I.e., only a small proportion of the

persons have survived in to the older age group. This type of population is found in primitive societies and is sometimes known as expansive (type I).

2) Transitional:- characterized by high birth rate and reduced death rate, with high (rapid) growth rate ("young population"). The drop in the death rate is usually brought about by improved medical care following socio –economic development. The shape of the population pyramid is triangular characterizing a developing society (sometimes known as expansive (type II).

3) Post –transitional:- characterized by low birth and death rates with stable, moderate growth rate. Narrow based pyramid and steeper sides. This typical of advanced or developed countries and is sometimes known as stationary (Type III). Life expectancy is higher and a high proportion of the population survives in to the old age (" old population").

Population Pyramid

The age –sex distribution of a population can be most clearly presented in a graphical form known as a "**population pyramid**". It is the most widely used of all graphic devices in population studies.

Population pyramids present the population of an area or country in terms of its composition by age and sex at a point in time. By convention, males are shown on the left of the pyramid, females on the right, young persons at the bottom, and the elderly at the top.

The pyramid consists of a series of bars, each drawn proportionately to represent the percentage contribution of each age-sex group (often in five-year groupings) to the total population; that is, the total area of the bars represents 100% of the population.

A pyramid conveys at a glance the entire shape of the age structure. It shows any gross irregularities due to special past events (such as a war, epidemic or age-selective migration), fluctuations of fertility, etc.. We refer to a population as “old” or “young”, according to the relative weight of old and young age groups in the total.

Population pyramid can be constructed either using absolute number or percentages. When constructing percent population pyramids, take the percentage from the group total.

The following pyramid is drawn for data obtained from census conducted in Cheha District, Gurage Zone, July 2000.

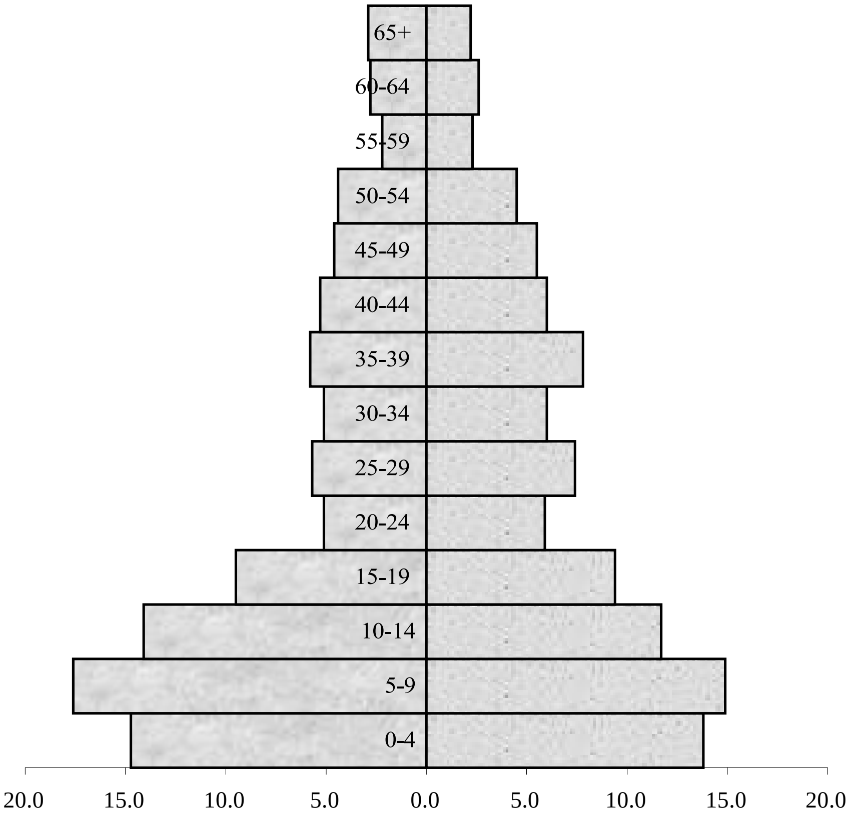


Fig. 11: Population Pyramid of Cheha District, Gurage Zone, July 2000

4.5 RATIO, PROPORTION AND RATE

4. **Ratio:** A ratio quantifies the magnitude of one occurrence or condition in relation to another.

1. **Sex Ratio** (SR): sex ratio is defined as the total number of male population per 100 female population,

SR = $\frac{M}{F}$ x 100 where M and F are total number of male and female populations, respectively.

Sex ratios are used for purposes of comparison.

a) The balance between the two sexes
b) The variation in the sex balance at different ages
c) It is also used for detecting errors in demographic data

2. **Child-Woman-Ratio** (CWR): It is defined as the ratio of the number of children under 5 years of age to the number of women in the childbearing age group (usually 15-49).

CWR = $P_{0\text{-}4} / P^{f}_{15\text{-}49}$ x 1000 = Number of children under 5 years of age per 1000 women in the child bearing age.

The child woman ratio is also known as measure of effective fertility because we are considering survivals up to the age of 4 not the dead ones.

3. **Dependency Ratio (DR)**: The dependency ratio describes the relation between the potentially self-supporting portion of the population and the dependent portions at the extremes of age. It is useful in economic studies.

$$\mathrm{DR} = \frac{P_{0\text{-}14} + P_{65^+}}{P_{15-64}} \times 100 = \frac{P_{0-14}}{P_{15-64}} \times 100 + \frac{P_{65^+}}{P_{15-64}} \times 100$$

= [Child dependency ratio] + [Age dependency ratio]

4.5.2 Proportion:- is a type of ratio which quantifies occurrences in relation to the population in which these occurrences take place. I.e., the numerator is also included in the denominator.

Example: The proportion of malaria cases among inhabitants of a certain locality.

4.5.3 Rate: A rate is a proportion with a time element, i.e., in which occurrences are quantified over a period of time.

The term rate appropriately refers to the ratio of demographic events to the population at risk in a specified period.

$$\text{Rate} = \frac{\text{Number of demographic events of interest}}{\text{Population at risk}} \times k$$

where K is a constant mainly a multiple of 10 (100, 1000, 10000, etc.).

Population at risk: This could be the mid-year population (population at the first of July 1), population at the beginning of the year or a more complex definition. Period for a rate is usually a year.

Rate could be crude or specific

It is considered as crude when it shows the frequency of a class of events through out the entire population without regarding to any of the smaller groupings. Crude rates are highly sensitive to the structure (age) of the population and are not directly used for comparison purposes. Where as a specific rate implies the events in a particular category of age, sex, race, particular disease, or other classification variables are used.

4.6 MEASURES OF FERTILITY

1. **Crude Birth Rate** (CBR): is the number of live births in a year per 1000 mid year population in the same year.

$$\text{CBR} = \frac{\text{Total Number of live birth}^{*} \text{ in a year}}{\text{Mid-year population in the same year}} \times 1000$$

(*) ***Live Birth*** *is the complete expulsion or extraction from its mother as a product of conception irrespective of the duration of pregnancy, which after such separation show evidence of life, (like breathing, pulsation of the heart, etc.).*

2. **General Fertility Rate** (GFR): is the number of births in a specified period per 1000 women aged 15-49 year; i.e.

$$\text{GFR} = \frac{\text{Total number of live births during a year}}{\text{Mid year female population aged 15-49 years}} \times 1000$$

GFR is general because we are considering all females in the age group 15-49 with out restricting to those who have child (children).

3. **Age Specific Fertility Rate** (ASFR): is the number of live births in a specified period per 1000 women of a given age or age group.

$$\text{ASFR} = \frac{\begin{array}{c}\text{Total number of live births registered to women}\\ \text{of a given age group during a year}\end{array}}{\begin{array}{c}\text{Mid year female population on the same}\\ \text{age group during the same year}\end{array}} \times 1000$$

ASFR is used to measure the reproductivity performance of a given age, thus showing variation in fertility by age.

4. **Total Fertility Rate** (TFR): is the sum of all age specific fertility rates for each year of age from 15-49 years. It is the average number of children that a synthetic (artificial) cohort (a group of persons who share a common experience within a defined period) of women would have at

the end of reproduction, if there were no mortality among women of reproductive age; each woman will live up to 49 years of age, about a total of 35 years.

$$\mathrm{TFR} = \sum_{i=15}^{49} \frac{B_i}{P_i^f} \times 1000 \text{ for single year classification of age}$$

$$\mathrm{TFR} = 5 \times \sum_{i=1}^{7} \frac{B_i}{P_i^f} \times 1000 \text{ for 5 year age group classification}$$

where $B_i = B_f + B_m$ = birth of both sex at age i of mothers

P_i^f = female population at age (age interval) i.

TFR is used as a standardized index for the over all fertility level and measures the total number of children a cohort of women will have at the end of their reproductive age.

5. **Gross Reproduction Rate** (GRR): is the total fertility rate restricted to female births only. Here, in order to compute the rate we need the ASFRs restricted to female births at each single year of age.

$$\mathrm{GRR} = \frac{B^f}{B^t} \sum_{i=15}^{49} \frac{B_i}{P_i^f} \times 1000 \text{ for single year age classification}$$

$$\mathrm{GRR} = \frac{B^f}{B^t} \times 5 \sum_{i=1}^{7} \frac{B_i}{P_i^f} \times 1000 \text{ for 5 years age grouping}$$

where $B^t = B^f + B^m$ (Total male and female births)

The above two formulae are used if we don't have female birth and female population by age. But if we have female birth and female population by age

$$\mathrm{GRR} = \sum_{i=15}^{49} \frac{B_i^f}{P_i^f} \times 1000 \text{ for single year age classification}$$

$$\mathrm{GRR} = 5 \times \sum_{i=1}^{7} \frac{B_i^f}{P_i^f} \times 1000 \text{ for 5 years age grouping}$$

GRR gives the average number of daughters a synthetic cohort (group) of women would have at the end of reproduction, in the absence of mortality.

Example: If GRR = 1000 => the current generation of females of child bearing age will maintain itself on the basis of current fertility rate with out mortality.

If GRR > 1000 => no amount of reduction of deaths will enable it to escape decline sooner or later and if GRR < 1000 the reverse happens.

In the absence of birth data cross classified by age of mother at birth and sex of the new born, we can approximate GRR from TFR simply by multiplying TFR by proportion of female births on the assumption that sex ratio at birth is constant. That is, the ratio of the number of male births to the number of female births remains constant over all ages of mothers.

$$\mathrm{GRR} = \frac{B^f}{B^t}\mathrm{TFR} = \frac{\mathrm{TFR}}{1+\frac{\text{Total male births}}{\text{Total female births}}} = \frac{\mathrm{TFR}}{1+\frac{\text{Sex ratio at birth}}{100}}$$

Note: A rate is Birth Rate if the denominator is mid year population and it is fertility rate if the denominator is restricted to females in the child bearing age.

6. Net reproduction rate (NRR)

The main disadvantage of the gross reproduction rate is that it does not take into account the fact that not all the females will live until the end of the reproductive period. In computing the net reproductionrate, mortality of the females is taken into account. The net reproduction rate measures the extent to which the females in the childbearing age-groups are replacing themselves in the next generation. The net reproduction rate is one in a stationarypopulation; a population which neither increases nor decreases (i.e. $r = 0$). In most cases, NRR is expressed per woman instead of per 1000 women.

NRR = 1 ⇒ stationary population (i.e., 1 daughter per woman)

NRR < 1⇒ declining population

NRR > 1⇒ growing population

4.7 MEASURES OF MORTALITY

1. **Crude Death Rate (CDR)**: is defined as total number of deaths due to all causes occurring in a defined area during a defined period per 1000 mid year population in the same area during the same period.

$$\mathrm{CDR} = \frac{\begin{array}{c}\text{Total number of deaths due to all causes}\\ \text{occurring in a an area in a given year}\end{array}}{\begin{array}{c}\text{Mid year population in the same area}\\ \text{in the given year}\end{array}} \times 1000$$

where mid year population is population of the area as of July 1 (middle of the year).

CDR measures the rate at which deaths are taking place from all causes in a given population during a specified year.

2. **Age-Specific Death Rate** (ASDR): is defined as total number of deaths occurring in a specified age group of the population of a defined area during a specified period per 1000 mid year population of the same age group of the same area during the same period.

$$\mathrm{ASDR}_a = \frac{\text{Total deaths at age or age group a}}{\text{Mid year population at age or age group a}} \times 1000$$

3. **Cause Specific Death Ratio and Rate:** A cause specific death ratio (proportionate mortality ratio) represents the percent of all deaths due to a particular cause or group of causes.

CSD ratio for cause c = $\frac{D_c}{D_t} \times 1000$, where Dc is total deaths from cause c and D_t is total deaths from all causes in a specified time period.

Cause Specific Death Rate (CSDR) is the number of deaths form cause c during a year per 1000 of the mid year population, i.e.

$$\mathrm{CSDR_c} = \frac{\text{Total deaths from a given cause c}}{\text{Population at risk}} \times 1000$$

4. **Infant Mortality Rate** (IMR): measures the risk of dying during infancy (i.e. the first age of life), and is defined as:

$$\mathrm{IMR} = \frac{\text{Deaths of children under one year of age}}{\text{Total live births}} \times 1000$$

Infant Mortality rate: the probability of dying between birth and age one year per 1000 live births.

5. **Neonatal Mortality Rate** (NMR): measures the risk of dying within 28 days of birth. It is defined as

$$\text{NMR} = \frac{\text{Deaths of children under 28 days of age}}{\text{Total live Births}} \times 1000$$

5. **Post - Neonatal Mortality Rate** (PNMR): Measures the risk of dying during infancy after the first 4 weeks of life, and is defined as:

$$\text{PNMR} = \frac{\text{Deaths of children aged 28 days to under one year}}{\text{Total live births}} \times 100$$

6. **Maternal Mortality Rate** (MMR): is defined as the number of deaths of mothers (Dm) due to maternal causes, i.e. complications of pregnancy, child birth, and puerperium, per 100,000 live births during a year, i.e.

$$\text{MMR} = \frac{\begin{array}{c}\text{Deaths of Mothers due to} \\ \text{maternal causes in a year}\end{array}}{\text{Total live births in the same year}} \times 100{,}000$$

MMR measures the risk of dying of mothers from maternal causes. Ideally the denominator should include all deliveries and abortions.

4.8 POPULATION GROWTH AND PROJECTION

The rate of increase or decline of the size of a population by natural causes (births and deaths) can be estimated crudely by using the measures related to births and deaths in the following way:

Rate of population growth

Crude Birth Rate - Crude Death Rate = **crude rate of natural increase.** This rate is based on naturally occurring events – births and deaths. When the net effect of migration is added to the natural increase it gives what is known as total increase.

Based on the total rate of increase (r), the population (P_t) of an area with current population size of (P_o) can be projected at some time t in the short time interval (mostly not more than 5 years) using thefollowing formula.

$P_t = P_o(1+r)^t$ **OR** $P_t = P_o \times \text{Exp}(r \times t)$ - the exponential

projection formula

For example if the CBR=46, CDR=18 per 1000 population and population size of 25,460 in 1998, then

Crude rate of natural increase = 46 - 18 = 28 per 1000 = 2.8 percent per year. The net effect of migration is assumed to be zero.

The estimated population in 2003, after 5 years, using the first formula will be

$P_{2003} = P_{1998} (1 + 0.028)^t = 25{,}460(1 + 0.028)^5 = 25{,}460(1.028)^5$

$= 25{,}460(1.148) = 29{,}230$

The population of the area in 2003 will be about 29,230.

Population doubling time

The doubling time of the size of a population can be estimated based on the formula for projecting the population.

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

From the above formula, the time at which the current population P_o will be $2 \times P_o$ can be found by:

$$2\times P_o = P_o(1+r)^t \Rightarrow 2 = (1+r)^t \Rightarrow \log 2 = t \times \log(1+r) \Rightarrow t = \frac{\log 2}{\log(1+r)}$$

For example, for the above community, r=0.028, then the doubling time for this population will be:

$$t = \frac{\log(2)}{\log(1+0.028)} = \frac{0.30103}{0.01199} = 25.1 \text{ years}$$

Therefore, it will take 25.1 years for the size of this population to be doubled.

A more practical approach to calculate the population doubling time is:

$2\times Po = Po\,(1+r)^t \Rightarrow 2 = (1+r)^t = (e^r)^t = (e)^{rt}$ (provided r is very small compared to 1)

$\Rightarrow \ln 2 = rt$

$\Rightarrow \ln 2 = rt$

$\Rightarrow 0.693 \approx 0.7 = rt$

Hence, $t = \dfrac{0.7}{r}$

For the above example, the doubling time(t) would be (0.7 / 0.028) = 25 years.

4.9 HEALTH SERVICES STATISTICS

Health Service 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.

1) What is the leading cause of death in the area? Is it malaria, tuberculosis, etc?
2) At what age is the mortality highest, and from what disease?
3) 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)
4) In comparison with similar areas, is this area healthier or not?
5) Are the health institutions in the area able to cope with the disease problem?
6) Is there any season at which various diseases have a tendency to break out? If so, can these be distinguished?
7) What are the factors involved in the incidence of certain diseases, like malaria, tuberculosis, etc.?

Uses of Health Statistics

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

1) Describe the level of community health
2) Diagnose community ills
3) Discover solutions to health problems and find clues for

administrative action

4) Determine priorities for health programmes
5) Develop procedures, definitions, techniques such as recording systems, sampling schemes, etc.
6) Promote health legislation
7) Create administrative standards of health activities
8) Determine the met and unmet health needs
9) Disseminate information on the health situation and health programmes
10) Determine success or failure of specific health programmes or undertake overall evaluation of public health work
11) Demand public support for health work

Major limitations of morbidity and mortality data from health institutions in Ethiopia include the following:

1) **Lack of completeness**: Health services at present (in 2000) cover only 47% of the population
2) **Lack of representativeness:** Illnesses and deaths recorded by health institutions do not constitute a representative sample of all illnesses & deaths occurring in the community.

3) **Lack of denominator**: The underlying population served by a health institution is difficult to define
4) **Lack of uniformity in quality**: No laboratory facilities in health

stations. Such facilities are available in hospitals.

5) **Lack of compliance with reporting:** Reports may be incomplete, not sent on time or not sent at all.

Health service utilization rates (Hospital statistics) - Indices relating to the hospital

1) **Admission rate (AR)**: The number of (hospital) admissions per 1000 of the population per year

$$\text{AR} = \frac{\text{Number of Admissions in the year}}{\text{Total Population of the Catchment area}} \times 1000$$

“Admission” is the acceptance of an in-patient by a hospital.

Discharges and deaths: The annual number of discharges includes the number of patients who have left the hospital (cured, improved, etc.), the number who have transferred to another health institution, and the number who have died.

2) **Average length of stay (ALS):** This index indicates the average period in hospital (in days) per patient admitted. Ideally, this figure should be calculated as:

$$\text{ALS} = \frac{\text{The Annual Number of Hospitalized Patient Days}}{\text{Number of Discharges and Deaths}}$$

That is, cumulative number of bed-days of all discharged patients (including those dying in hospital) during one year divided by the number of discharged and dead patients in the same year.

3) **Bed-occupancy rate (BOR**): This figure expresses the average percentage occupancy of hospital beds.

$$\text{BOR} = \frac{\text{The Annual Number of Hospitalized Patient Days}}{\text{Total Number of Beds}} \times \frac{1}{365}$$

4) **Turnover interval (TI):** the turnover interval expresses the average period, in days, that a bed remains empty, in other words, the average time elapsing between the discharge of one patient and the admission of the next.

$$\text{TI} = \frac{(365 \times \text{Number of Beds}) \text{ - Number of Hospitalized Patient Days}}{\text{Number of Discharges and Deaths}}$$

This figure is obtained by subtracting the actual number of hospitalization days from the potential number of hospitalization days in a year and dividing the result by the number of discharges (and deaths) in the same year.

The turnover interval is zero when the bed-occupancy rate is 100%.

very short or negative turnover interval points to a shortage of beds, whereas a long interval may indicate an excess of beds or a defective admission mechanism.

5) Hospital Death Rate (HDR)

$$\text{HDR} = \frac{\text{Total Number of Hospital Deaths in a Given Period}}{\text{Number of Discharges in the Given Period}} \times 1000$$

4.10 Exercises

a) Calculate the population doubling time of a given country with annual rate of growth (r) = 1%.

b) The following summary table was taken from the annual (1988) health profile of district z.

Year	Total population of the district	No of health institutions in the district			Total number of hospital beds
		Health Station	Health Center	Hospital	
1988	400,000	14	2	1	80

During the same year, there were 14,308 discharges and deaths. The annual number of hospitalized patient days was also recorded as 28,616.

i) Calculate:

1. the health service coverage of the district
2. .the average length of stay
3. the bed occupancy rate
4. the turnover interval

ii) What do you understand from your answers in parts 1 and 4?

iii) Show that the average time that elapsed between the discharge of one patient and the admission of the next was about **one** hour.

CHAPTER FIVE

ELEMENTARY PROBABILITY AND PROBABILITY DISTRIBUTIONS

5.1 Learning Objectives

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

1. Understand the concepts and characteristics of probabilities and probability distributions
2. Compute probabilities of events and conditional probabilities
3. Differentiate between the binomial and normal distributions
4. Understand the concepts and uses of the standard normal distribution

5.2 INTRODUCTION

In general, there is no completely satisfactory definition of probability. Probability is one of those elusive concepts that virtually everyone knows but which is nearly impossible to define entirely adequately.

A fair coin has been tossed 800 times. Here is a record of the number of times it came up head, and the proportion of heads in the throws already made:

Number of tosses	Number of heads in the last 80 tosses	Cumulative number of heads	Proportion of heads in total number of tosses
80	38	38	.475
160	40	78	.488
240	47	125	.521
320	39	164	.513
400	40	204	.510
480	45	249	.519
560	32	281	.502
640	40	321	.502
720	38	359	.499
800	42	401	.501

As the result of this experiment, we say that the probability of heads [notation: pr(H)] on one toss of this coin is about .50. That is, Pr(H) = .5

Definition: The probability that something occurs is the proportion of times it occurs when exactly the same experiment is repeated a very large (preferably infinite!) number of times in independent trials, “independent” means the outcome of one trial of the experiment doesn’t affect any other outcome. The above definition is called the **frequentist definition of Probability.**

The Classical Probability Concept

If there are n ***equally likely possibilities***, of which one must occur and m are regarded as favourable, or as a “success,” then the probability of a “success” is m/n.

Example: What is the probability of rolling a 6 with a *well-balanced die*?

In this case, m=1 and n=6, so that the probability is 1/6 = 0.167

Definitions of some terms commonly encountered in probability

Experiment : In statistics anything that results in a count or a measurement is called an experiment. It may be the parasite counts of malaria patients entering Felege Hiwot Hospital , or measurements of social awareness among mentally disturbed children or measurements of blood pressure among a group of students.

Sample space: The set of all possible outcomes of an experiment , for example, (H,T).

Event: Any subset of the sample space H or T.

5.3 Mutually exclusive events and the additive law

Two events A and B are mutually exclusive if they have no elements in common. If A and B are outcomes of an experiment they cannot both happen at the same time. That is, the occurrence of A precludes the occurrence of B and vice versa. For example, in the toss of a coin, the event A (it lands heads) and event B (it lands tails) are mutually exclusive. In the throw of a pair of dice, the event A (thesum of faces is 7) and B (the sum of faces is 11) are mutually exclusive.

The additive law, when applied to two mutually exclusive events, states that the probability of either of the two events occurring is obtained by adding the probabilities of each event. Thus, if A and B are mutually exclusive events,

Pr(A or B) = Pr (A) + Pr(B).

Extension of the additive law to more than two events indicates that if A, B, C… are mutually exclusive events, **Pr(A or B or C or…) = Pr (A) + pr(B)+ pr(C) + …**

Eg. One die is rolled. Sample space = S = (1,2,3,4,5,6)

Let A = the event an odd number turns up, A = (1,3,5)

Let B = the event a 1,2 or 3 turns up; B = (1,2,3)

Let C = the event a 2 turns up, C= (2)

i) Find Pr (A); Pr (B) and Pr (C)

Pr(A) = Pr(1) + Pr(3) + Pr(5) = 1/6+1/6+ 1/6 = 3/6 = 1/2

Pr(B) = Pr(1) + pr(2) + Pr(3) = 1/6+1/6+1/6 = 3/6 = ½

Pr (C) = Pr(2) = 1/6

ii) Are A and B; A and C; B and C mutually exclusive?

- A and B are not mutually exclusive. Because they have the elements 1 and 3 in common

- similarly, B and C are not mutually exclusive. They have the element 2 in common.

- A and C are mutually exclusive. They don't have any element in common

When A and B are not mutually exclusive pr(A or B) = Pr(A) + Pr(B) cannot be used. The reason is that in such a situation A and B overlap in a venn diagram, and the elements in the overlap are counted twice. Therefore, when A and B are not mutually exclusive**,** Pr(A or B) = Pr (A)

+ Pr(B) – Pr(A and B). The formula considered earlier for mutually exclusive events is a special case of this, since pr(A and B) = 0.

Eg. Of 200 seniors at a certain college, 98 are women, 34 are majoring in Biology, and 20 Biology majors are women. If one student is chosen at random from the senior class, what is the probability that the choice will be either a Biology major or a women).

Pr (Biology major or woman) = Pr (Biology major) + Pr(woman) - Pr (Biology major and woman) =34/200 + 98/200 - 20/200 = 112/200 = .56

5.4 Conditional probabilities and the multiplicative law

Sometimes the chance a particular event happens depends on the outcome of some other event. This applies obviously with many events that are spread out in time.

Eg. The chance a patient with some disease survives the next year depends on his having survived to the present time. Such probabilities are called conditional.

The notation is Pr(B/A), which is read as "the probability event B occurs given that event A has already occurred ."

Let A and B be two events of a sample space S. The conditional probability of an event A, given B, denoted by Pr (A/B)= P(A n B) / P(B) , $P(B) \neq 0$.

Similarly, P(B/A) = P(A n B) / P(A) , $P(A) \neq 0$. This can be taken as an alternative form of the multiplicative law.

Eg. Suppose in country X the chance that an infant lives to age 25 is .95, whereas the chance that he lives to age 65 is .65. For the latter, it is understood that to survive to age 65 means to survive both from birth to age 25 and from age 25 to 65. What is the chance that a person 25 years of age survives to age 65?

a

Notation	Event	Probability
A	Survive birth to age 25	.95
A and B	Survive both birth to age 25 and age 25 to 65	.65B/A
	Survive age 25 to 65 given survival to age 25	?

Then, Pr(B/A) = Pr(A n B) / Pr(A) = .65/.95 = .684 . That is, a person aged 25 has a 68.4 percent chance of living to age 65.

Independent Events

Often there are two events such that the occurrence or nonoccurrence of one does not in any way affect the occurrence or nonoccurrence of the other. This defines independent events. Thus, if events A and B are independent, Pr(B/A) = P(B); Pr(A/B) = P(A).

Eg. 1) A classic example is n tosses of a coin and the chances that on each toss it lands heads. These are independent events. The chance of heads on any one toss is independent of the number of previous heads. No matter how many heads have already been observed, the chance of heads on the next toss is ½.

Eg 2) A similar situation prevails with the sex of offspring. The chance of a male is approximately ½. Regardless of the sexes of previous offspring, the chance the next child is a male is still ½.

with independent events, the multiplicative law becomes :

$$\Pr(A \text{ and } B) = \Pr(A)\Pr(B)$$

Hence, $\Pr(A) = \Pr(A \text{ and } B) / \Pr(B)$, where $\Pr(B) \neq 0$

$\Pr(B) = \Pr(A \text{ and } B) / \Pr(A)$, where $\Pr(A) \neq 0$

EXERCISE

Consider the drawing of two cards one after the other from a deck of 52 cards. What is the probability that both cards will be spades?

a) with replacement
b) without replacement

Summary of basic Properties of probability

1. Probabilities are real numbers on the interval from 0 to 1; i.e., $0 \le Pr(A) \le 1$
2. If an event is certain to occur, its probability is 1, and if the event is certain not to occur, its probability is 0.
3. If two events are mutually exclusive (disjoint), the probability that one or the other will occur equals the sum of the probabilities; Pr(A or B) = Pr(A) + Pr(B).
4. If A and B are two events, not necessarily disjoint, then Pr(A or B)= Pr (A) +Pr (B) – Pr(A and B).
5. The sum of the probabilities that an event will occur and that it will not occur is equal to 1; hence, $P(A') = 1 - P(A)$
6. If A and B are two independent events, then Pr (A and B) = Pr (A) Pr (B)

5.5 Random variables and probability distributions

Usually numbers can be associated with the outcomes of an experiment. For example, the number of heads that come up when a coin is tossed four times is 0, 1,2,3 or 4. Sometimes, we may find a situation where the elements of a sample space are categories. In such cases, we can assign numbers to the categories .

Eg. There are 2,500 men and 2000 women in a senior class. Assume a person is randomly selected .

Sample space	Number assigned
Man	1
Woman	2

Pr (Man) = .56 That is, Pr (1) = .56

Pr (Women) = .44 That is, Pr (2) = .44

Definition: A random variable for which there exists a discrete definition of values with specified probabilities is a discrete random variable.

Definition: A random variable whose values form a continuum (i.e., have no gaps) such that ranges of values occur with specified probabilities is a continuous random variable.

The values taken by a discrete random variable and its associated probabilities can be expressed by a rule, or relationship that is called a *probability mass (density) function.*

Definition: A *probability distribution (mass function*)is a mathematical relationship, or rule, that assigns to any possible value of a discrete random variable X the probability P(X = xi). This assignment is made for all values xi that have positive probability. The probability distribution can be displayed in the form of a table giving the values and their associated probabilities and/or it can be expressed as a mathematical formula giving the probability of allpossible values.

General rules which apply to any probability distribution:

1. Since the values of a probability distribution are probabilities, they must be numbers in the interval from 0 to 1.
2. Since a random variable has to take on one of its values, the sum of all the values of a probability distribution must be equal to 1.

Eg. Toss a coin 3 times. Let x be the number of heads obtained. Find the probability distribution of x .

f (x) = Pr (X = xi) , i = 0, 1, 2, 3.
Pr (x = 0) = 1/8.. TTT
Pr (x = 1) = 3/8 ……………………………. HTT THT TTH
Pr (x = 2) = 3/8 ……………………………..HHT THH HTH Pr
(x = 3) = 1/8 .... HHH

Probability distribution of X.

X = xi	0	1	2	3
Pr(X=xi)	1/8	3/8	3/8	1/8

The required conditions are also satisfied. i) $f(x) \geq 0$ ii) $\sum f(x_i) = 1$

5.5.1 THE EXPECTED VALUE OF A DISCRETE RANDOM VARIABLE

The expected value, denoted by E(x) or μ, represents the "average" value of the random variable. It is obtained by multiplying eachpossible value by its respective probability and summing over all the values that have positive probability.

Definition: The expected value of a discrete random variable is defined as

$$E(X) = \mu = \sum_{i=1}^{n} x_i P(X = x_i)$$

Where the x_i's are the values the random variable assumes with positive probability

Example: Consider the random variable representing the number of episodes of diarrhoea in the first 2 years of life. Suppose this random variable has a probability mass function as below

R	0	1	2	3	4	5	6
P(X = r)	.129	.264	.271	.185	.095	.039	.017

What is the expected number of episodes of diarrhoea in the first 2 years of life?

E(X) = 0(.129)+1(.264)+2(.271)+3(.185)+4(.095)+5(.039)+6(.017)= 2.038

Thus, on the average a child would be expected to have 2 episodes of diarrhoea in the first 2 years of life.

i. THE VARIANCE OF A DISCRETE RANDOM VARIABLE

The variance represents the spread of all values that have positive probability relative to the expected value. In particular, the variance is obtained by multiplying the squared distance of each possible value from the expected value by its respective probability and summing overall the values that have positive probability.

Definition: The variance of a discrete random variable denoted by X is defined by

$$V(X) = \sigma^2 = \sum_{i=1}^{k} (x_i - \mu)^2 P(X = x_i) = \sum_{i=1}^{k} x_i^2 P(X = x_i) - \mu^2$$

Where the X_i's are the values for which the random variable takes on positive probability. The SD of a random variable X, denoted by SD(X) or σ is defined by square root of its variance.

Example: Compute the variance and SD for the random variable representing number of episodes of diarrhoea in the first 2 years of life.

$E(X) = \mu = 2.04$

$\sum_{i=1}^{n} x_i P(X = x_i)$ = $0^2(.129)$ + $1^2(.264)$ + $2^2(.271)$ + $3^2(.185)$ +

$4^2(.095) + 5^2(.039) + 6^2(0.017) = 6.12$

Thus, $V(X) = 6.12 - (2.04)^2 = 1.967$ and the SD of X is

$$\sigma = \sqrt{1.967} = 1.402$$

ii. THE BINOMIAL DISTRIBUTION

Binomial assumptions:

1) The same experiment is carried out n times (n trials are made).

2) Each trial has two possible outcomes (usually these outcomes are called “ success” and “ failure”. Note that a successful outcome does not imply a good one, nor failure a bad outcome. If P is the probability of success in one trial, then , 1-p is the probability of failure.

3) The result of each trial is independent of the result of any other trial.

Definition: If the binomial assumptions are satisfied, the probability of r successes in n trials is:

$$P(X = r) = \binom{n}{r} P^{r} (1 - P)^{n-r}$$, r = 0, 1, 2, …, n

This probability distribution is called the binomial distribution.

$\binom{n}{r}$ = ${}_{n}C_{r}$ is the number of ways of choosing r items from n, and is a number we have to calculate. The general formula for the coefficients $\binom{n}{r}$ is $\binom{n}{r} = \frac{n!}{r!(n-r)!}$

If the true proportion of events of interest is P, then in a sample of size n the mean of the binomial distribution is **n×p** and the standard deviation is $\sqrt{np(1-p)}$

Example: Assume that, when a child is born, the probability it is a girl is ½ and that the sex of the child does not depend on the sex of an older sibling.

A) Find the probability distribution for the number of girls in a family with 4 children.

B) Find the mean and the standard deviation of this distribution.

$f(x) = p(X=r) = {}_4C_r \ (1/2)^r (1/2)^{4-r}$; X= 0 ,1, 2, 3, 4.

A) Probability distribution

X	0	1	2	3	4
P(X=r)	1/16	4/16	6/16	4/16	1/16

B) Mean = nP = 4 x 1/2 = 2

Standard deviation = $\sqrt{nP(1-P)}$ = $\sqrt{4\times1/2\times1/2}$ = $\sqrt{1}$ = 1

Exercise: Suppose that in a certain malarious area past experience indicates that the probability of a person with a high fever will be positive for malaria is 0.7. Consider 3 randomly selected patients (with high fever) in that same area.

1) What is the probability that no patient will be positive for malaria?
2) What is the probability that exactly one patient will be positive for malaria?
3) What is the probability that exactly two of the patients will be positive for malaria?
4) What is the probability that all patients will be positive for malaria?
5) Find the mean and the SD of the probability distribution given above.

Answer: 1) 0.027 2) 0.189 3) 0.441 4) 0.343

5) $\mu = 2.1$ and $\sigma = 0.794$

5.5.4 The Normal Distribution

The **Normal Distribution** is by far the most important probability distribution in statistics. It is also sometimes known as the Gaussian distribution, after the mathematician Gauss. The distributions of many medical measurements in populations follow a normal distribution (eg. Serum uric acid levels, cholesterol levels, blood pressure, height and weight). The normal distribution is a theoretical, continuous probability distribution whose equation is:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for } -\infty < x < +\infty$$

The area that represents the probability between two points c and d on abscissa is defined by:

$$P(c < X < d) = \int_c^d \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

The important characteristics of the Normal Distribution are:

1) It is a probability distribution of a continuous variable. It extends from minus infinity(-∞) to plus infinity (+∞).
2) It is unimodal, bell-shaped and symmetrical about x = u.
3) It is determined by two quantities: its mean **(** μ) and SD (σ). Changing μ alone shifts the entire normal curve to the left or right. Changing σ alone changes the degree to which the distribution is spread out.
4. The height of the frequency curve, which is called the probability density, cannot be taken as the probability of a particular value. This is because for a continuous variable there are infinitely many

possible values so that the probability of any specific value is zero.

5. An observation from a normal distribution can be related to a standard normal distribution (**SND**) which has a published table. Since the values of μ and σ will depend on the particular problem in hand and tables of the normal distribution cannot be published for all values of μ and σ, calculations are made by referring to the standard normal distribution which has μ **= 0** and σ **= 1.** Thus an observation x from a normal distribution with mean μ and standard deviation σ can be related to a Standard normal distribution by calculating :

   SND = Z = (x - μ **) /** σ

Area under any Normal curve

To find the area under a normal curve (with mean μ and standard deviation σ) between x=a and x=b, find the Z scores corresponding to a and b (call them Z_1 and Z_2) and then find the area under the standard normal curve between Z_1 and Z_2 from the published table.

Z- Scores

Assume a distribution has a mean of 70 and a standard deviation of 10.

How many standard deviation units above the mean is a score of 80?

$$(80-70) / 10 = 1$$

How many standard deviation units above the mean is a score of 83?

$$Z = (83 - 70) / 10 = 1.3$$

The number of standard deviation units is called a Z-score or Z-value.

In general, Z = (raw score - population mean) / population SD **= $(x-\mu) / \sigma$**

In the above population, what Z-score corresponds to a raw score 68?

$$Z = (68-70)/10 = - 0.2$$

Z-scores are important because given a Z – value we can find out the probability of obtaining a score this large or larger (or this low or lower). (look up the value in a z-table). To look up the probability of obtaining a Z-value as large or larger than a given value, look up the first two digits of the Z-score in the left hand column and then read the hundredths place across the top.

Hence, $P(-1 < Z < +1) = 0.6827$; $P(-1.96 < Z < +1.96) = 0.95$ and $P(-2.576 < Z < + 2.576) = 0.99$.

From the symmetry properties of the stated normal distribution,

$P(Z \le -x) = P(Z \ge x) = 1 - P(z \le x)$

Example1: Suppose a borderline hypertensive is defined as a person whose DBP is between 90 and 95 mm Hg inclusive, and the subjects are 35-44-year-old males whose BP is normally distributed with mean 80 and variance 144. What is the probability that a randomly selected person from this population will be a borderline hypertensive?

Solution: Let X be DBP, X ~ N(80, 144)

$$P(90 < X < 95) = P\left(\frac{90-80}{12} < \frac{x-\mu}{\sigma} < \frac{95-80}{12}\right) = P(0.83 < z < 1.25)$$

$$= P(Z < 1.25) - P(Z < 0.83) = 0.8944 - 0.7967 = 0.098$$

Thus, approximately 9.8% of this population will be borderline hypertensive.

Example2: Suppose that total carbohydrate intake in 12-14 year old males is normally distributed with mean 124 g/1000 cal and SD 20 g/1000 cal.

a) What percent of boys in this age range have carbohydrate intake above 140g/1000 cal?

b) What percent of boys in this age range have carbohydrate intake below 90g/1000 cal?

Solution: Let X be carbohydrate intake in 12-14-year-old males and

$X \sim N(124, 400)$

a) $P(X > 140) = P(Z > (140-124)/20) = P(Z > 0.8)$

$= 1 - P(Z < 0.8) = 1 - 0.7881 = 0.2119$

b) $P(X < 90) = P(Z < (90-124)/20) = P(Z < -1.7)$

$= P(Z > 1.7) = 1 - P(Z < 1.7) = 1 - 0.9554 = 0.0446$

b. Exercises

1. Assume that among diabetics the fasting blood level of glucose is approximately normally distribute with a mean of 105 mg per 100 ml and SD of 9 mg per 100 ml.

a) What proportions of diabetics have levels between 90 and 125 mg per 100 ml?

b) What proportions of diabetics have levels below 87.4 mg per 100 ml?

c) What level cuts of the lower 10% of diabetics?

d) hat are the two levels which encompass 95% of diabetics?

Answers a) 0.9393b) 0.025 c) 93.48 mg per 100 ml

d) X_1 = 87.36 mg per 100 ml and X_2 = 122.64 mg per 100 ml

2. Among a large group of coronary patients it is found that their serum cholesterol levels approximate a normal distribution. It was found that 10% of the group had cholesterol levels below 182.3 mg per 100 ml where as 5% had values above 359.0 mg per 100 ml.What is the mean and SD of the distribution?

Answers: mean = 260 ml per 100 ml and standard deviation = 60 mg per 100 ml

3. Answer the following questions by referring to the table of the standard normal distribution.

a) If Z = 0.00, the area to the right of Z is______.
b) If Z = 0.10, the area to the right of Z is______.

c) If Z = 0.10, the area to the left of Z is______.

d) If Z = 1.14, the area to the right of Z is______.

e) If Z = -1.14, the area to the left of Z is_______.

If Z = 1.96, the area to the right of Z is_______and the area to the left of Z = - 1.96 is_________. Thus, the central 95% of the standard normal distribution lies between –1.96 and 1.96 with ___% in each tail.

CHAPTER SIX

SAMPLING METHODS

6.1 LEARNING OBJECTIVES

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

1. Define population and sample and understand the different sampling terminologies
2. Differentiate between probability and Non-Probability sampling methods and apply different techniques of sampling
3. Understand the importance of a representative sample
4. Differentiate between random error and bias
5. Enumerate advantages and limitations of the different sampling methods

6.2 INTRODUCTION

Sampling involves the selection of a number of a study units from a defined population. The population is too large for us to consider collecting information from all its members. If the whole population is taken there is no need of statistical inference. Usually, a representative subgroup of the population (sample) is included in the investigation. A representative sample has all the important characteristics of the population from which it is drawn.

Advantages of samples

- cost - sampling saves time, labour and money
  - quality of data - more time and effort can be spent on getting reliable data on each individual included in the sample.
  - Due to the use of better trained personnel, more careful supervision and processing a sample can actually produce precise results.

If we have to draw a sample, we will be confronted with the following questions:

a) What is the group of people (population) 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?

Apart from persons, a population may consist of mosquitoes, villages, institutions, etc.

6.3 Common terms used in sampling

Reference population (also called source population or target population) - the population 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.

Study or sample population - the population included in the sample.

Sampling unit - the unit of selection in the sampling process

Study unit - the unit on which information is collected.

- the sampling unit is not necessarily the same as the study unit.
- if the objective is to determine the availability of latrine, then the study unit would be the household; if the objective is to determine the prevalence of trachoma, then the study unit would be the individual.

Sampling frame - the list of all the units in the reference population, from which a sample is to be picked.

Sampling fraction (Sampling interval) - the ratio of the number of units in the sample to the number of units in the reference population **(n/N)**

6.4 Sampling methods (Two broad divisions)

6.4.1 Non-probability Sampling Methods

- Used when a sampling frame does not exist
- No random selection (unrepresentative of the given population)
- Inappropriate if the aim is to measure variables and generalize findings obtained from a sample to the population.

Two such non-probability sampling methods are:

A) 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.

B) Quota sampling: is a method that ensures that a certain number of sample units from different categories with specific 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.

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

6.4.2 Probability Sampling methods

- A sampling frame exists or can be compiled.
- Involve random selection procedures. All units of the population should have an equal or at least a known chance of being included in the sample.
- Generalization is possible (from sample to population)

A) Simple random sampling (SRS)

- This is the most basic scheme of random sampling.
- Each unit in the sampling frame has an equal chance of being selected
- representativeness of the sample is ensured.

However, it is costly to conduct SRS. Moreover, minority subgroups of interest in the population my not be present in the sample in sufficient numbers for study.

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 study units, using a “lottery” method or a table of random numbers.

"**Lottery” method**: for a small population it may be possible to use the “lottery” method: each unit in the population is represented by a slip of paper, these are put in a box and mixed, and a sample of the required size is drawn from the box.

Table of random numbers: if there are many units, however, the above technique soon becomes laborious. Selection of the units is greatly facilitated and made more accurate by using a set of random numbers in which a large number of digits is set out in random order. The property of a table of random numbers is that, whichever way it is read, vertically in columns or horizontally in rows, the order of the digits is random. Nowadays, any scientific calculator has the same facilities.

B) Systematic Sampling

Individuals are chosen at regular intervals (for example, every k^{th}) from the sampling frame. The first unit to be selected is taken at random from among the first k units. For example, a systematic sample is to be selected from 1200 students of a school. The sample size is decided to be 100. The sampling fraction is: 100 /1200 = 1/12. Hence, the sample interval is 12.

The number of the first student to be included in the sample is chosen randomly, for example by blindly picking one out of twelve pieces of paper, numbered 1 to 12. If number 6 is picked, every twelfth student will be included in the sample, starting with student number 6, until 100 students are selected. The numbers selected would be 6,18,30,42,etc.

Merits

- Systematic sampling is usually less time consuming and easier to perform than simple random sampling. It provides a good approximation to SRS.

- Unlike SRS, systematic sampling can be conducted without a sampling frame (useful in some situations where a sampling frame is not readily available).

Eg., In patients attending a health center, where it is not possible to predict in advance who will be attending.

Demerits

- If there is any sort of cyclic pattern in the ordering of the subjects which coincides with the sampling interval, the sample will not be representative of the population.

Examples

- List of married couples arranged with men's names alternatively with the women's names (every 2nd, 4th , etc.) will result in a sample of all men or women).
- If we want to select a random sample of a certain day (sampling fraction on which to count clinic attendance, this day may fall on the same day of the week, which might, for example be a market day.

C) Stratified Sampling

It is appropriate when the distribution of the characteristic to be studied is strongly affected by certain variable (heterogeneous population). The population is first divided into groups (strata) according to a characteristic of interest (eg., sex, geographic area, prevalence of disease, etc.). A separate sample is then taken independently from each stratum, by simple random or systematic sampling.

- **proportional allocation** - if the same sampling fraction is used for each stratum.

- **non- proportional allocation** - if a different sampling fraction is used for each stratum or if the strata are unequal in size and a fixed number of units is selected from each stratum.

Merit

- The representativeness of the sample is improved. That is, adequate representation of minority subgroups of interest can be ensured by stratification and by varying the sampling fraction between strata as required.

DEMERIT

- Sampling frame for the entire population has to be prepared separately for each stratum.

D) Cluster sampling

In this sampling scheme, selection of the required sample is done on groups of study units (clusters) instead of each study unit individually. The sampling unit is a cluster, and the sampling frame is a list of these clusters.

procedure

- The reference population (homogeneous) is divided into clusters. These clusters are often geographic units (eg districts, villages, etc.)
- A sample of such clusters is selected
- All the units in the selected clusters are studied

Ethiopia Public Health Training Initiative • Ethiopia Public Health Training Initiative • EPHTI

It is preferable to select a large number of small clusters rather than a small number of large clusters.

Merit

A list of all the individual study units in the reference population is not required. It is sufficient to have a list of clusters.

Demerit

It is based on the assumption that the characteristic to be studied is uniformly distributed throughout the reference population, which may not always be the case. Hence, sampling error is usually higher than for a simple random sample of the same size.

E) Multi-stage sampling

This method is appropriate when the reference population is large and widely scattered . Selection is done in stages until the final sampling unit (eg., households or persons) are arrived at. The primary sampling unit (PSU) is the sampling unit (usually large size) in the first sampling stage. The secondary sampling unit (SSU) is the sampling unit in the second sampling stage, etc.

Example - The PSUs could be *kebeles* and the SSUs could be households.

Merit - Cuts the cost of preparing sampling frame

Demerit - Sampling error is increased compared with a simple random sample.

Multistage sampling gives less precise estimates than sample random sampling for the same sample size, but the reduction in cost usually far outweighs this, and allows for a larger sample size.

6.5 Errors in sampling

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.

6.5.1 Sampling error (random error)

A sample is a subset of a population. Because of this property of samples, results obtained from them cannot reflect the full range of variation found in the larger group (population). This type of error, arising from the sampling process itself, is called sampling error,160 which is a form of random error. Sampling error can be minimized by increasing the size of the sample. When n = N $\Rightarrow$ sampling error = 0

6.5.2 Non-sampling error (bias)

It is a type of systematic error in the design or conduct of a sampling procedure which results in distortion of the sample, so that it is no longer representative of the reference population. We can eliminate or reduce the non-sampling error (bias) by careful design of the sampling procedure and not by increasing the sample size.

Example: If you take male students only from a student dormitory in Ethiopia in order to determine the proportion of smokers, you would result in an overestimate, since females are less likely to smoke. Increasing the number of male students would not remove the bias.

- There are several possible sources of bias in sampling (eg., accessibility bias, volunteer bias, etc.)
- The best known source of bias is non response. It is the failure to obtain information on some of the subjects included in the sample to be studied.
- Non response results in significant bias when the following two conditions are both fulfilled.

  - When non-respondents constitute a significant proportion of the sample (about 15% or more)
  - When non-respondents differ significantly from respondents.

- There are several ways to deal with this problem and reduce the possibility of bias:

  a) Data collection tools (questionnaire) have to be pre-tested.

  b) If non response is due to absence of the subjects, repeated attempts should be considered to contact study subjects who were absent at the time of the initial visit.

  c) To include additional people in the sample, so that non-respondents who were absent during data collection can be replaced (make sure that their absence is not related to the topic being studied).

NB: 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.

CHAPTER SEVEN

ESTIMATION

7.1 Learning objectives

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

1. Understand the concepts of sample statistics and population parameters
2. Understand the principles of sampling distributions of means and proportions and calculate their standard errors
3. Understand the principles of estimation and differentiate between point and interval estimations
4. Compute appropriate confidence intervals for population means and proportions and interpret the findings
5. Describe methods of sample size calculation for cross – sectional studies

7.2 Introduction

In this chapter the concepts of sample statistics and population parameters are described. The sample from a population is used to provide the estimates of the population parameters. The standard error, one of the most important concepts in statistical inference, is introduced. Methods for calculating confidence intervals for population means and proportions are given. The importance of the normal distribution (Z distribution) is stressed throughout the chapter.

7.3 Point Estimation

Definition: A parameter is a numerical descriptive measure of a population (μ is an example of a parameter). A statistic is a numerical descriptive measure of a sample ($\overline{X}$ is an example of a statistic).
To each sample statistic there corresponds a population parameter. We use $\overline{X}$, S^2, S , p, etc. to estimate μ, σ^2, σ, P (or π), etc.

Sample statistic	**Corresponding population parameter**
$\overline{X}$ (sample mean)	μ (population mean)
S^2 (sample variance)	σ^2 (population variance)
S (sample Standard deviation)	σ(population standard deviation)
p (sample proportion)	P or π (Population proportion)

We have already seen that the mean $\overline{X}$ of a sample can be used to estimate μ.This does not, of course, indicate that the mean of every sample will equal the population mean.

Definition: A point estimate of some population parameter O is a single value Ô of a sample statistic.

Eg. The mean survival time of 91 laboratory rats after removal of the thyroid gland was 82 days with a standard deviation of 10 days (assume the rats were randomly selected).

In the above example, the point estimates for the population parameters μ and σ (with regard to the survival time of all laboratory rats after removal of the thyroid gland) are 82 days and 10 days respectively.

7.4 Sampling Distribution of Means

The sampling distribution of means is one of the most fundamental concepts of statistical inference, and it has remarkable properties. Since it is a frequency distribution it has its own mean and standard deviation .

One may generate the sampling distribution of means as follows:

1) Obtain a sample of n observations selected completely at random from a large population . Determine their mean and then replace the observations in the population.
2) Obtain another random sample of n observations from the population, determine their mean and again replace the observations.
3) Repeat the sampling procedure indefinitely, calculating the mean of the random sample of n each time and subsequently replacing the

observations in the population.

4) The result is a series of means of samples of size n. If each mean in the series is now treated as an individual observation and arrayed in a frequency distribution, one determines the sampling distribution of means of samples of size n.

Because the scores ($\overline{X}$ s) in the sampling distribution of means are themselves means (of individual samples), we shall use the notation $\sigma\,\overline{X}$ for the standard deviation of the distribution. The standard deviation of the sampling distribution of means is called the standard error of the mean.

Eg.
- Obtain repeat samples of 25 from a large population of males.
- Determine the mean serum uric acid level in each sample by replacing the 25 observations each time.
- Array the means into a distribution.
- Then you will generate the sampling distribution of mean serum uric acid levels of samples of size 25.

Properties

1. The mean of the sampling distribution of means is the same as the population mean, μ .

2. The SD of the sampling distribution of means is $\sigma / \sqrt{n}$.

3. The shape of the sampling distribution of means is approximately a normal curve, regardless of the shape of the population distribution and provided n is large enough (Central limit theorem).

In practice, the approximation is a workable one if n is 30 or more.

Eg 1. Suppose you have a population having four members with values 10,20,30 and 40 . If you take all conceivable samples of size 2 with replacement:

a) What is the frequency distribution of the sample means ?

b) Find the mean and standard deviation of the distribution (standard error of the mean).

Possible samples	x i (sample mean)
(10, 20) or (20, 10)	15
(10, 30) or (30, 10)	20
(10, 40) or (40, 10)	25
(20, 30) or (30, 20)	25
(20, 40) or (40, 20)	30
(30, 40) or (40, 30)	35
(10, 10)	10
(20, 20)	20
(30, 30)	30
(40, 40)	40

a) frequency distribution of sample means

sample mean $(\bar{x}i)$	**frequency (fi)**
10	1
15	2
20	3
25	4
30	3
35	2
40	1

b) i) The mean of the sampling distribution = $\sum x \text{ifi} / \sum \text{fi}$

$= 400 / 16 = 25$

ii) The standard deviation of the mean $= \sigma \bar{x} = \sqrt{\sum (\bar{x}\text{i} - \mu)^2 / \sum \text{fi}}$

$= \sqrt{\{\sum (10 - 25)^2 + (15 - 25)^2 + \ldots + (40 - 25)^2\} / 16}$

$= \sqrt{1000 / 16} = \sqrt{62.5} = 7.9$

Eg2. For the population given above (10,20,30 and 40)

f) Find the population mean. Show that the population mean (μ) = the mean of the sampling distribution

g) Find the population standard deviation and show that the standard error of the mean $(\sigma \bar{X} = \sigma_x / \sqrt{n})$

(that is, the standard error of the mean is equal to the population standard deviation divided by the square root of the sample size)

Answers to example 2

a) $\mu = \sum xi / N = (10 + 20 + 30 + 40) / 4 = 25$

b) $\sigma^2 = \sum(xi - \mu)^2 / N = (225+ 25+ 25 + 225) / 4 = 125$

Hence, $\sigma = \sqrt{125} = 11.18$ and σ_x (standard error) $=\sigma_x / \sqrt{n} = 11.180 / 1.414 = 7.9$

7.5 Interval Estimation (large samples)

A point estimate does not give any indication on how far away the parameter lies. A more useful method of estimation is to compute an interval which has a high probability of containing the parameter.

Definition: An interval estimate is a statement that a population parameter has a value lying between two specified limits.

7.5.1 Confidence interval for a single mean

Consider the standard normal distribution and the statement $\Pr(-1.96 \leq Z \leq 1.96) = .95$

This is merely a shorthand algebraic statement that 95% of the standard normal curve lies between + 1.96 and –1.96. If one chooses the sampling distribution of means (a normal curve with mean μ and standard deviation $\sigma/\sqrt{n}$), then ,

$$\Pr(-1.96 \le (\bar{X} - \mu)/(\sigma/\sqrt{n}) \le 1.96) = .95$$

A little manipulation without altering the probability value of 95 percent gives

$$\Pr(\bar{X} - 1.96(\sigma/\sqrt{n}) \le \mu \le \bar{X} + 1.96(\sigma/\sqrt{n})) = .95$$

The range **$\bar{X} - 1.96(\sigma/\sqrt{n})$** to **$\bar{X} + 1.96(\sigma/\sqrt{n})$)** is called the 95% confidence interval;

$\bar{X} - 1.96(\sigma/\sqrt{n})$ is the lower confidence limit while **$\bar{X} + 1.96(\sigma/\sqrt{n})$** is the upper confidence limit.

The confidence, expressed as a proportion, that the interval **$\bar{X} - 1.96(\sigma/\sqrt{n})$ to $\bar{X} + 1.96(\sigma/\sqrt{n})$** contains the unknown population mean is called the confidence coefficient. When this coefficient is .95 as given above, the following formal definition of confidence interval is given. If many different random samples are taken, and if the confidence interval for each is determined, then it is expected that 95% of these computed intervals will contain the population mean (μ) .

Clearly, there appears to be no rationale (logical basis) for taking repeated samples of size n and determine the corresponding confidence intervals. However, the knowledge of the properties of these sampling distributions of means (if one hypothetically obtained these repeated samples) permits one to draw a conclusion based upon one sample and this was shown repeatedly in the previous sections.

From the above definition of Confidence interval (C.I.), the widely used definition is derived. That is, when one claims $\overline{X}$ ± **1.96 (σ/$\sqrt{}$n)** as the limits on μ, there is a 95% chance that the statement is correct (that μ is contained within the interval).

If more than 95% certainty regarding the population mean - say, a 99% C.I. were desired, the only change needed is to use ±2.58 (the point enclosing 99% of the standard normal curve), which gives $\overline{X}$ ± **2.58 (σ/$\sqrt{}$n) .**

Eg 1. The mean reading speed of a random sample of 81 adults is 325 words per minute. Find a 90% C.I. For the mean reading speed of all adults (μ) if it is known that the standard deviation for all adults is 45 words per minute.

Given

n = 81

σ = 45

x = 325

Z = ± 1.64 (the point enclosing 90% of the standard normal curve)

A 90% C.I. for μ is $\bar{x}$ ± **1.64 (σ /√n) = 325 ± (1.64 x 5) = 325 ± 8.2**
= (316.8, 333.2)

Therefore, A 90% CI. For μ is 316.8 to 333.2words per minute.

Eg 2. A random sample of 100 drug-treated patients has a mean survival time of 46.9 months. If the SD of the population is 43.3 months, find a 95% confidence interval for the population mean.

(The population consists of survival times of cancer patients who have been treated with a new drug)

46.9 ± (1.96) (43.3 /√100) = 46.9 ± 8.5 = (38.4 to 55.4 months)

Hence, there is 95% certainty that the limits (38.4 , 55.4) embrace the mean survival times in the population from which the sample arose.

7.5.2 Confidence interval for the difference of means

Consider two different populations. The first population (X) has mean μ_x and standard deviation σ_x, the second (Y) has mean μ_y and standard deviation σ_y. From the first population take a sample of size n_x and compute its mean $\bar{x}$; from the second population take independently a sample of size n_y and compute $\bar{y}$; then determine $\bar{x} - \bar{y}$. Do this for all pairs of samples that can be chosen independently

from the two populations. The differences, $\bar{x} - \bar{y}$, are a new set of scores which form the sampling distribution of differences of means.

The characteristics of the sampling distribution of differences of means are:

1) The mean of the sampling distribution of differences of means equals the difference of the population means (Mean = $\mu_x - \mu_y$).

2) The standard deviation of the sampling distribution of differences of means, also called the standard error of differences of means is denoted by $\sigma_{(\bar{x} - \bar{y})}$.

$\sigma_{(\bar{x} - \bar{y})} = \sqrt{(\sigma^2_{\bar{x}} + \sigma^2_{\bar{y}})}$ where $\sigma_{\bar{x}}$ is the standard error of the mean of the first population and $\sigma_{\bar{y}}$ is the standard error of the mean of the second population. ($\sigma^2_{\bar{x}} = \sigma^2_x / n_x$; $\sigma^2_{\bar{y}} =$; σ^2_y / n_y)

3) The sampling distribution is normal if both populations are normal, and is approximately normal if the samples are large enough (even if the populations aren't normal). In practice, it is assumed that the sampling distribution of differences of means is normal if both n_x and n_y are ≥ 30.

A formula for C.I is found by solving $Z = \{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)\} / \sigma_{(\bar{x} - \bar{y})}$ for $\mu_x - \mu_y$; hence C.I. for the difference of means is $(\bar{x} - \bar{y}) \pm Z.\sigma_{(\bar{x} - \bar{y})}$

Eg1. If a random sample of 50 non-smokers have a mean life of 76 years with a standard deviation of 8 years, and a random sample of 65 smokers live 68 years with a standard deviation of 9 years,

A) What is the point estimate for the difference of the population means?

B) Find a 95% C.I. for the difference of mean lifetime of non-smokers and smokers.

Given

Population x(non-smokers) n_x=50 , $\bar{x} = 76$, $S_x = 8$, $\sigma^2 \bar{x} = S^2 / n_{x,x} = 8^2/50$ =1.28 years

Population y (smokers) $n_y=65$, $\bar{y} = 68$, $S_y = 9$, $\sigma^2 \bar{y} = S^2 / n_y$, = $9^2/65$ =1.25 years

A) A point estimate for the difference of population means **($\mu_x - \mu_y$) =** $\bar{x} - \bar{y}$ **=76-68 = 8 years**

B) At a 95% confidence level, Z = ± 1.96, $\sigma(\bar{x} - \bar{y}) = \sqrt{1.28 + 1.25}$

$= \sqrt{2.53} = 1.59$ years

Hence, 95% C.I. for $\mu_x - \mu_y$ $= (\bar{x} - \bar{y}) \pm 1.96\, \sigma(\bar{x} - \bar{y}) = 8 \pm 1.96\ (1.59)$

$= 8 \pm 3.12$ = (4.88 to 11.12 years)

Exercise

An anthropologist who wanted to study the heights of adult men and women took a random sample of 128 adult men and 100 adult women and found the following summary results.

	Mean height	**Standard deviation**
Adult men	170 cms	8 cma
Adult women	164 cms	6 cms

Find a 95% C.I for the difference of mean height of adult men and women.

7.5.3 Confidence interval for a single proportion

Notation: P (or π) = proportion of “successes” in a population (parameter)

Q = 1-P = proportion of “failures” in a population

p = proportion of successes in a sample

q = 1-p proportion of “failures” in a sample

σ_p= Standard deviation of the sampling distribution of proportions

= Standard error of proportions

n = size of the sample

The population represents categorical data while the scores in the sampling distribution are proportions between 0 and 1.This set of proportions has a mean and standard deviation. The sampling distribution of proportions has the following characteristics:

1. Its mean = P, the proportion in the population.

2. $\sigma_p = \sqrt{PQ/n}$

3 The shape is approximately normal provided n is sufficiently large -in this case, $nP \geq 5$ and $nQ \geq 5$ are the requirements for sufficiently large n (central limit theorem for proportions) .

The confidence interval for the population proportion (P) is given by the formula:

$p \pm Z\,\sigma_p$ (that is, $p - Z\,\sigma_p$ and $p + Z\,\sigma_p$)

p = sample proportion, σ_p = standard error of the proportion (= $\sqrt{PQ/n}$).

Example: An epidemiologist is worried about the ever increasingtrend of malaria in a certain locality and wants to estimate the proportion of persons infected in the peak malaria transmission period.If he takes a random sample of 150 persons in that locality during the peak transmission period and finds that 60 of them are positive for malaria, find

a) 95% b)90% c)99% confidence intervals for the proportion ofthe whole infected people in that locality during the peak malaria transmission period .

Sample proportion = 60 / 150 = .4

The standard error of proportion depends on the population P. However, the population proportion (P) is unknown. In such situations, $\sqrt{pq/n}$ can be used as an approximation to σ_p = $\sqrt{(.4x.6)/150}$ = .04

a) A 95% C.I for the population proportion (the proportion of the whole infected people in that locality) = .4 ± 1.96 (.04) = (.4 ± .078) = (.322, .478).

b) A 90% C.I for the population proportion (the proportion of the whole infected people in that locality) = .4 ± 1.64 (.04) = (.4 ± .066) = (.334, .466).

A 99% C.I for the population proportion (the proportion of the whole infected people in that locality) = .4 ± 2.58 (.04) = (.4 ± .103) = (.297, .503).

7.5.4 Confidence interval for the difference of two proportions

By the same analogy, the C.I. for the difference of proportions (P_x - P_y) is given by the following formula.

C.I. for $\mathbf{P_x - P_y = (p_x - p_y) \pm Z\, \sigma_{(Px - Py)}}$. Where Z is determined by the confidence coefficient and $\sigma_{(Px - Py)} = \sqrt{\{(p_x q_x)/n_x + (p_x q_x)/n_x\}}$

Example: Each of two groups consists of 100 patients who have leukaemia. A new drug is given to the first group but not to the second (the control group). It is found that in the first group 75 people have remission for 2 years; but only 60 in the second group. Find 95%

confidence limits for the difference in the proportion of all patients with leukaemia who have remission for 2 years.

Note that $n_xp_x = 100 \times .75 = 75 > 5$

$n_xq_y = 100 \times .25 = 25 > 5$

$n_yp_y = 100 \times .60 = 60 > 5$

$n_yq_y = 100 \times .40 = 40 > 5$

$p_x = .75$, $q_x = .25$, $n_x = 100$, $\sigma^2_{Px} = p_xq_x / n_x = .75 \times .25 / 100 = .001875$

$p_y = .60$, $q_x = .40$, $n_y = 100$, $\sigma^2_{Py} = p_yq_y / n_y = .60 \times .40 / 100 = .0024$

Hence, $\sigma^2_{(Px-Py)} = \sqrt{(\sigma^2_{Px} + \sigma^2_{Px})} = \sqrt{(p_xq_x / n_x) + (p_yq_y / n_y)} = \sqrt{.001875+.0024} = .065$

At a 95% Confidence level, Z = ± 1.96 and the difference of the two independent random samples is (.75 - .60) = .15 . Therefore, a 95 % C. I. for the difference in the proportion with 2-year remission is (.15 ± 1.96 (.065)) = (.15 ± .13) = (.02 to .28).

7.6 Sample Size Estimation in cross – sectional studies

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 analyse it.

7.6.1 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 ± 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 ($N \geq 10{,}000$)
  is:

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

Show how the above formula is obtained.

A 95% C.I. for $P = p \pm 1.96$ se , if we want our confidence interval to have a maximum width of $\pm w$,

$1.96 \text{ se} = w$

$1.96 \sqrt{p(1-p)/n} = w$

$1.96^2 \, p(1-p)/n = w^2$, Hence, $n = 1.96^2 \, p(1-p)/w^2$

Example 1

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 .

$821.25 / (1+ (821.25/3000)) = 644.7 \approx 645$ subjects

7.6.2 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:

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

Eg. A health officer wishes to estimate mean haemoglobin level in a defined community. From preliminary contact he thinks this mean is about 150 mg/l with a standard deviation of 32 m/l. If he is willing to tolerate a sampling error of up to 5 mg/l in his estimate, how many subjects should be included in his study? (α =5%, two sided)

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

$n = (1.96)^2 (32)^2 / (5)^2 = 157.4 \approx 158$ persons

- If the population size is , say, 2000 ,
  The required sample size would be 146 persons.

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

7.6.3 Comparison of two Proportions (sample size in each region)

$n = (p_1q_1 + p_2q_2)\,(f(\alpha,\beta)) / ((p_1 - p_2)$

α = type I error (level of significance)

β = type II error ($1-\beta$ = power of the study)

power = the probability of getting a significant result

$f(\alpha,\beta) = 10.5$, 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

7.6.4 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.

Eg. The birth weights in districts A and B will be compared. In district A the mean birth weight is expected to be 3000 grams with a standard deviation of 500 grams. In district B the mean is expected to be 3200 grams with a standard deviation of 500 grams. The required sample size to demonstrate (with a likelihood of 90% , that is with a power of 90%) a significant difference between the mean birth weights in districts A and B would be:

$N = (1.96 + 1.28)^2 (500 + 500)^2 / (3200 – 3000)^2$

= 131 newborn babies in each district
Note that $f(\alpha,\beta) = 10.5$
That is , $\alpha = .05$ (two sided) $\Rightarrow Z = 1.96$
$\beta = (1- .9) = .1$ (one sided) $\Rightarrow Z = 1.28$

7.7 Exercises

1. Of 45 patients treated by a 1 hour hypnosis session to kick the smoking habit, 36 stopped smoking, at least for the moment. Find a 95% C.I. for the proportion of all smokers who quit after choosing this type of treatment. (the patients were selected randomly).

   A 95% confidence interval for the population proportion (i.e., proportion of all smokers who quit smoking after choosing hypnosis) is (0.68 to 0.92).

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?

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

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

CHAPTER EIGHT

HYPOTHESIS TESTING

8.1 Learning objectives

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

1. Understand the concepts of null and alternative hypothesis
2. Explain the meaning and application of statistical significance
3. Differentiate between type I and type II errors
4. Describe the different types of statistical tests used when samples are large and small
5. Explain the meaning and application of P – values
6. Understand the concepts of degrees of freedom

8.2 Introduction

In chapter 7 we dealt with estimation which is one form of statistical inference. In this chapter we shall introduce a different form of inference, the significance test or hypothesis test. A significance test enables us to measure the strength of the evidence which the data supply concerning some proposition of interest.

Definition :A statistical hypothesis is an assumption or a statement which may or may not be true concerning one or more populations.

Eg. 1) The mean height of the Gondar College of Medical Sciences (GCMS) students is 1.63m.

2) There is no difference between the distribution of Pf and Pv malaria in Ethiopia (are distributed in equal proportions.)

In general, hypothesis testing in statistics involves the following steps:

1. Choose the hypothesis that is to be questioned.
2. Choose an alternative hypothesis which is accepted if the original hypothesis is rejected.
3. Choose a rule for making a decision about when to reject the original hypothesis and when to fail to reject it.
4. Choose a random sample from the appropriate population and compute appropriate statistics: that is, mean, variance and so on.
5. Make the decision.

8.3 The null and alternative hypotheses

The main hypothesis which we wish to test is called the null hypothesis, since acceptance of it commonly implies “no effect” or “ no difference.” It is denoted by the symbol H_O .

H_O is always a statement about a parameter (mean, proportion, etc. of a population). It is not about a sample, nor are sample statistics used in formulating the null hypothesis. H_O is an equality ($\mu = 14$) rather than an inequality ($\mu \geq 14$ or $\mu < 14$).

Examples 1) H_O : $\mu = 1.63$ m (from the previous example).

2) At present only 60% of patients with leukaemia survive more than 6 years. A doctor develops a new drug. Of 40 patients, chosen at random, on whom the new drug is tested, 26 are alive after 6 years. Is the new drug better than the former treatment?

Here, we are questioning whether the proportion of patients who recover under the new treatment is still .60 (and hope that it will be improved; this will be shown in our choice of H_A in the next section).

The null hypothesis of the above statement is written as : $\mathbf{H_O : P = .60}$

Choosing the Alternative Hypothesis (H_A)

The notation H_A (or H1) is used for the hypothesis that will be accepted if H_O is rejected. H_A must also be formulated before a sample is tested, so it, like the null hypothesis (H_O), does not depend on sample values. If the mean height of the GCMS students ($H_O : \mu = 1.63$ m) is questioned, then the alternative hypothesis (H_A) is set $\mu \neq 163$ m. Other alternatives are also:

$H_A : \mu > 1.63$ m.

$H_A : \mu < 1.63$ m.

Possible choices of H_A

If H_O is	**then H_A is**
$\mu = A$ (single mean)	$\mu \neq A$ or $\mu < A$ or $\mu > A$
$P = B$ (single proportion)	$P \neq B$ or $P < B$ or $P > B$
$\mu_x - \mu_y = C$ (difference of means)	$\mu_x - \mu_y \neq C$ or $\mu_x - \mu_y < C$ or $\mu_x - \mu_y > C$
$P_x - P_y = D$(difference of proportions)	$P_x - P_y \neq D$ or $P_x - P_y < D$ or $P_x - P_y > D$

Where, A, B, C and D are constants.

Consider the previous example (patients with leukaemia)

$$H_O : P = .60$$
$$H_A : P > .60$$

The doctor is trying to reach a decision on whether to make further tests on the new drug. If the proportion of patients who live at least 6 years is not increased under the new treatment or is increased only by an amount due to sampling fluctuation, he will look for another drug. But if the proportion who are aided is significantly larger (that is, if he is able to conclude that the population proportion is greater than .60) - then he will continue his tests.

Exercises

State H_O and H_A for each of the following

1) Is the average height of the GCMS students 1.63 m or is it more?
2) Is the average height of the GCMS students 1.63 m or is it less?
3) Is the average height of the GCMS students 1.63 m or is it something different?
4) There is a belief that 10% of the smokers develop lung cancer in country x.
5) Are men and women infected with malaria in equal proportions, or is a higher proportion of men get malaria in Ethiopia?

8.4 Level of significance

A method for making a decision must be agreed upon. If H_O is rejected, then H_A is accepted. How is a “significant” difference defined? A null hypothesis is either true or false, and it is either rejected or not rejected. No error is made if it is true and we fail to reject it, or if it is false and rejected. An error is made, however, if it is true but rejected, or if it is false and we fail to reject it.

A random sample of size n is taken and the information from the sample is used to reject or accept (fail to reject) the null hypothesis. It is not always possible to make a correct decision since we are dealing with random samples. Therefore, we must learn to live with probabilities of type I (α) and type II (β) errors.

Definitions:

A Type I error is made when H_O is true but rejected.

A Type II error is made when H_O is false but we fail to reject it .

Notation: α is the probability of a type I error. It is called the **level of significance.**

β is the probability of a type II error.

The following table summarises the relationships between the null hypothesis and the decision taken .

Null hypothesis	Decision	
	Accept H_O (Fail to reject H_O)	Reject HO
H_O true	Correct	Type I error
H_O false	Type II error	Correct

In practice, the level of significance (α) is chosen arbitrarily and the limits for accepting H_O are determined. If a sample statistic is outside those limits, H_O is rejected (and H_A is accepted). The form of H_A will determine the kind of limits to be set up (either one tailed or two tailed tests) .

Consider the situation when H_A includes the symbol " $\neq$ " . That is, H_A: $\mu \neq \ldots$, $P \neq \ldots$, $\mu_x - \mu_y \neq \ldots$., $P_x - P_y \neq \ldots$. etc (two tailed test)

1. α (level of significance) is arbitrarily chosen, equal to a small number (usually .01, .05, etc..)
2. Z values are determined so that the area in each of the tails of the normal distribution is $\alpha / _2$.

The most common values of z are:

α	**.10**	**.05**	**.01**
Z	± **1.64**	± **1.96**	± **2.58**

3. The experiment is carried out and the Z value of the appropriate sample statistic (x, p, x-y, p_x - p_y) is determined. If the computed Z value falls within the limits determined in step 2 above, we fail to reject H_O; if the computed Z value is outside those limits, H_O is rejected (and H_A is accepted). Since they separate the “fail to reject” and “ reject” regions, the limits determined in step 2 will be referred to as the critical values of Z.

8.5 Tests of Significance on means and Proportions (large samples)

It is important to remember that a test of significance always refers to a null hypothesis. The concern here is with an unknown population parameter, and the null hypothesis states that it is some particular value.

The test of significance answers the question: Is chance (sampling) variation a likely explanation of the discrepancy between a sample result and the corresponding null hypothesis population value? A “yes” answer – a discrepancy that is likely to occur by chance variation– indicates the sample result is compatible with the claim that the sampling is from a population in which the null hypothesis prevails. This is the meaning of “not statistically significant.” A “no” answer – a discrepancy that is unlikely to occur by chance variation – indicates that the sample result is not compatible with the claim that sampling is from a population in which the null hypothesis prevails. This is the meaning of “statistically significant.”

As shown earlier, the level significance selected, be it 5 percent, 1 percent, or otherwise, must be clearly indicated. A statement that the results were “statistically significant” without giving further details is worthless.

P – Values

P – values abound in medical and public health research papers, so it is essential to understand precisely what they mean.

Having set up the null hypothesis, we then evaluate the probability that we could have obtained the observed data (or data that were more extreme) if the null hypothesis were true. This probability is usually called the P – value. If it is small, conventionally less than 0.05, the null hypothesis is rejected as implausible. In other words, an outcome that could occur less than one time in 20 when the null hypothesis is true would lead to the rejection of the null hypothesis. In this formulation, when we reject the null hypothesis we accept a complementary alternative hypothesis. If $P > 0.05$ this is often taken as suggesting that insufficient information is available to discount the null hypothesis.

When P is below the cut off level(α), say 0.05, the result is called **statistically significant**(and below some lower level, such as 0.01, it may be called **highly significant**); when above 0.05 it is called **not significant**. It is important to distinguish between the significance level and the p – value. The significance level α is the probability of making a type I error. This is set before the test is carried out. The P – value is the result observed after the study is completed and is based on the observed data.

It would be better (informative) to give the exact values of P; such as, $P = 0.02$ or $P = 0.15$ rather than $P < 0.05$ or $P > 0.05$. It is now increasingly common to see the expression of exact values largely due to the availability of computer programs which give the exact P – values.

8.5.1 Tests of significance on a single mean and comparison of two means

The preceding discussions provide the necessary equipment for conducting tests of significance on a single mean and comparison of two means.

A statistical test of significance on a single mean

One begins with a statement that claims a particular value for the unknown population mean. The statistical inference consists of drawing one of the following two conclusions regarding this statement:

I) Reject the claim about the population mean because there is sufficient evidence to doubt its validity.

II) Do not reject the claim about the population mean, because there is not sufficient evidence to doubt its validity.

The analysis consists of determining the chance of observing a mean as deviant as or more deviant than the sample mean, under the assumption that the sample came from a population whose mean is μ_O. One then compares this chance with the predetermined "sufficiently small" chance by referring to the table of the Z distribution (the standard normal distribution) . The critical ratio (Z statistic) is calculated as: $Z = (\bar{x} - \mu_O) / (\sigma / \sqrt{n})$.

Example: Assume that in a certain district the mean systolic blood pressure of persons aged 20 to 40 is 130 mm Hg with a standard deviation of 10 mm Hg . A random sample of 64 persons aged 20 to 40 from village x of the same district has a mean systolic blood pressure of 132 mm Hg. Does the mean systolic blood pressure of the dwellers of the village (aged 20 to 40) differ from that of the inhabitants of the district (aged 20 to 40) in general, at a 5% Level of significance?

H_O: $\mu = 130$ (the mean systolic blood pressure of the village
is the same as the mean SBP of the district)

H_A: $\mu \neq 130$

$\alpha = .05$ (that is, the probability of rejecting H_O when it is true is to be .05).

The area of each shaded "tail " of the standard normal curve is .025 and the corresponding Z scores (Z tabulated) at the boundaries are ± 1.96.

Sample: $n = 64$, $\overline{X} = 132$

The Z score for the random sample of 64 persons of the village aged 20 to 40 years:

$$\mathbf{Z\ calc = (132-130) / (10/ \sqrt{64}) = 2 / 1.25 = 1.6}$$

This score falls inside the “fail to reject region” from –1.96 to +1.96 .

If the calculated Z value is positive, the rule says:

reject H_O if Z calculated (Z calc) > Z tabulated (Z tab)
or accept H_O if Z calc < Z tab.

On the other hand, if the calculated Z value is negative:

reject H_O if Z calculated (Z calc) < Z tabulated (Z tab)

(Here, both Zcalculated and Ztabulated are negative values)

Hence, the null hypothesis of the above example is accepted. That is, the mean systolic blood pressure of persons (aged 20 to 40) living in village x is the same as the mean systolic blood pressure of the inhabitants (aged 20 to 40) of the district.

The same conclusion will be reached by referring to the corresponding P-value. We use table 4 to find the P-value associated with an observed (calculated) value of Z which is 1.6. From the table mentioned above we get P=.11 (note that this p-value is greater than the given level of significance).

Comparison of two Means

The purpose of this section is to extend the arguments of the single mean to the comparison of two sample means. Since medicine (public health) is, by nature, comparative, this is a rather widespread situation, more common than that of the single mean of the preceding section.

In the comparison of two means, there are two samples of observations from two underlying populations (often treatment and control groups) whose means are denoted by μ_t and μ_c and whose standard deviations are denoted by σ_t and σ_c . Recalling that a test of significance involves a null hypothesis that specifies values for population quantities, the relevant null hypothesis is that the means are identical, i.e.,

$$_O : \mu_t = \mu_c \quad \text{or} \quad H_O : \mu_t - \mu_c = 0$$

The rationale for the test of significance is as before. Assuming the null hypothesis is true (i.e., that there is no difference in the population means), one determines the chance of obtaining differences in sample means as discrepant as or more discrepant than that observed. If this chance is sufficiently small, there is reasonable evidence to doubt the validity of the null hypothesis; hence, one concludes there is a statistically significant difference between the means of the two populations (i.e., one rejects the null hypothesis).

Consider the Example Given in Section 7.5.2

Test the hypothesis that there is no difference between the mean lifetimes of on smokers and smokers at a .01 level of significance.

Population x (non-smokers) n_x=50 , $\bar{x}$ = 76, S_x = 8, $\sigma^2_{\bar{x}} = S^2_x / n_x = 8^2/50$ =1.28 years

Population y (smokers) n_y=65 , $\bar{y}$ = 68, S_y = 9, $\sigma^2_{\bar{y}} = S^2_y / n_y = 9^2/65$ =1.25 years

Hypotheses: $H_O : \mu_t = \mu_c$ **or** $H_O : \mu_t - \mu_c = 0$

$H_A : \mu_t \neq \mu_c$ **or** $H_A : \mu_t - \mu_c \neq 0$

α **= .01 (two tailed)** $\Rightarrow$ **Z (tabulated) =** $\pm$ **2.58**

Z calc = { $(\bar{x} - \bar{y}) - (\mu_x - \mu_y)$ **}/ Standard error of the difference of means**

Standard error of the difference of means = $\sigma_{(\bar{x} - \bar{y})} = \sqrt{1.28 + 1.25}$

= 1.59 years

Hence, Zcalc = (76 – 68) / 1.59 = 8 / 1.59 = 5.03

The corresponding P-value is less than .003.

Because Zcalc > Ztab (i.e., **P-value < the given α value),** the null hypothesis (H_O) is rejected. That is, there is a statistically significant difference in the mean lifetimes of nonsmkers and smokers

8.5.2 Test of significance on a single proportion and comparison of proportions

Test of Significance on a Single Proportion

Note that the expression given in section 7.5.3 applies in a similar way to this section.

Example: Among susceptible individuals exposed to a particular infectious agent, 36 percent generally develop clinical disease. Among a school group of 144 persons suspected of exposure to the agent, only 35 developed clinical disease. Is this result within chance variation (α = .05).

To use the table of the standard normal distribution, one needs to calculate critical ratios. There is, however, an additional consideration derived from the fact that the critical ratio is based on smooth normal curves used as substitutes for discrete binomial distributions. A correction continuity is therefore appropriate.

$$Z = \{|p - P| - (\frac{1}{2n})\} / se(P)$$

$$= \{|\frac{35}{144} - .36| - \frac{1}{288}\} / \sqrt{.36 \times .64/144} = 2.84$$

The observed value of Z (which is 2.84) corresponds to a P-value of .005.

Hence, at a .05 level of significance, the null hypothesis is rejected. Therefore, the sample finding is not compatible with the population proportion.

Comparison of two Proportions

A similar approach is adopted when performing a hypothesis test to compare two proportions. The standard error of the difference in proportions is again calculated, but because we are evaluating the probability of the data on the assumption that the null hypothesis is true we calculate a slightly different standard error. If the null hypothesis is true, the two samples come from populations having the same true proportion of individuals with the characteristic of interest, say, P . We do not know P, but both p_1 and p_2 are estimates of P. Our best estimate of P is given by calculating :

$$p = \frac{r1 + r2}{n1 + n2}$$

The standard error of $P_1 - P_2$ under the null hypothesis is thuscalculated on the assumption that the proportion in each group is p , sothat we have

$$se(P_1 - P_2) = \sqrt{p(1 - p)(1/n1 + 1/n2)}$$

The sampling distribution of $P_1 - P_2$ is normal, so we calculate a standard normal deviate (Z statistic) , as

$$Z = \frac{p1 - p2}{se(p1 - p2)}$$

Example: A health officer is trying to study the malaria situation of Ethiopia. From the records of seasonal blood survey (SBS) results he came to understand that the proportion of people having malaria in Ethiopia was 3.8% in 1978 (Eth. Cal). The size of the sample considered was 15000. He also realised that during the year that followed (1979), blood samples were taken from 10,000 randomly selected persons. The result of the 1979 seasonal blood survey showed that 200 persons were positive for malaria. Help the health officer in testing the hypothesis that the malaria situation of 1979 did not show any significant difference from that of 1978 (take the level of significance, α =.01).

$H_O : P_{1978} = P_{1979}$ (or $P_{1978} - P_{1979} = 0$)

$H_A : P_{1978} \neq P_{1979}$ (or $P_{1978} - P_{1979} \neq 0$)

$p_{1978} = .038$, $n_{1978} = 15,000$

$p_{1979} = .02$, $n_{1979} = 10,000$

Z tab ($\alpha = .01$) = 2.58

Common (pooled) proportion, $p = \dfrac{570 + 200}{15000 + 10000} = 0.0308$

and the standard error = $\sqrt{.038 \times .9692(1/15000 \ + 1/10000)}$

$= .0022$

Hence, Z calc = (.038 - .020) / .0022 = .018 / .0022 = 8. 2

Which corresponds to a P-value of less than .003.

Decision: reject H_o (because Z calc > Z tab); in other words, the p-value is less than the level of significance (i.e., $\alpha = .01$)

Therefore, it is concluded that there was a statistically significant difference in the proportion of malaria patients between 1978 and 1979 at a .01 level of significance.

Note that the effect of the continuity correction is always to reduce the magnitude of the numerator of the critical ratio. In this sense, the continuity correction is conservative in that its use, compared with its not being used, will label somewhat fewer observed sample differences as being statistically significant. On the other hand, when dealing with values of $n\pi$ and $n(1-\pi)$ that are well above 5, the continuity correction has a negligible effect. This negligible effect can be easily seen from the above example.

8.6 One tailed tests

In the preceding sections we have been dealing with two tailed (sided) tests.

consider the situation when H_A includes the symbol " > or < ". That is,

H_A: μ >__ , H_A: μ <__, H_A : P >__, H_A : P <__, H_A : $\mu_x - \mu_y$ >___, H_A : $\mu_x - \mu_y$ <___, etc. (One tailed test).

In the previous sections we learned how to set up a null hypothesis H_O and an alternative hypothesis H_A , and how to reach a decision on whether or not to reject the null hypothesis at a given level of significance when H_A includes the symbol " $\neq$ ". The decision was reached by following a specific convention (i.e., the area in each tail of the sampling distribution is assumed to be $\alpha/_2$) . This convention determines two values of Z which separate a " fail to reject region" from two rejection regions. Then the Z-value of a statistic is computed for a random sample, and H_O is still accepted if the Z-value falls in the " fail to reject region " previously determined; otherwise, H_O is rejected.

A one tailed test ($H_{A:}$ μ >_, etc.) is also justified when the investigator can state at the outset that it is entirely inconceivable that the true population mean is below (or above) that of the null hypothesis. To defend this position, there must be solid and convincing supporting evidence. In medical (public health) applications, however, a one tailed test is not commonly encountered.

When a one-tailed test is carried out, it is important to notice that the rejection region is concentrated in one tail of the sampling distribution of means. The area of this one tail is α, rather than the $\alpha/_2$ used previously, and the critical value of Z changes accordingly.

Example:

Two-tailed test	**one-tailed test**
$H_{O:}\ \mu = 100$	$H_{O}:\ \mu = 100$
$H_{O:}\ \mu \neq 100$	$H_{A:}\ \mu > 100$
$\alpha = .05$	$\alpha = .05$

The most frequently used values of α and the corresponding critical values of Z are:

α **(level of significance)**	**two-tailed**	**one -tailed, <**	**one-tailed, >**
.10	± **1.64**	**- 1.28**	**1.28**
.05	± **1.96**	**- 1.64**	**1.64**
.01	± **2.58**	**- 2.33**	**2.33**

Logically, an investigator chooses between a one and two-tailed test before he obtains his sample result, i.e., the choice is not influenced by the sample outcome. The investigator asks: Is it important to be able to detect alternatives only above the null hypothesis mean, or is it important to be able to detect alternatives that may be either above or below this mean? His answer, depends on the particular circumstances of the investigation.

Example1: It is known that 1-year old dogs have a mean gain in weight of 1.0 pound per month with a standard deviation of .40 pound. A special diet supplement is given to a random sample of 50 1-year old dogs for a month; their mean gain in weight per month is 1.15 pounds, with a standard deviation of .30 pound. Does weight gain in 1-year old dogs increase if a special diet supplement is included in their usual diet? (.01 level of significance)

$H_{O:}\ \mu = 1.0$

$H_{A:}\ \mu > 1.0$ (one tailed test)

Z tab (α =.01) = 2.33 and reject H_O if Z calc > 2.33.

Z calc = (1.15 - 1.0) / (.4/√50) = 0.15/0.0566 = 2.65 (This corresponds to a P-value of .004)

Hence, at a .01 level of significance weight gain is increased if a special diet supplement is included in the usual diet of 1-year old dogs.

Example 2: A pharmaceutical company claims that a drug which it manufactures relieves cold symptoms for a period of 10 hours in 90% of those who take it. In a random sample of 400 people with colds who take the drug, 350 find relief for 10 hours. At a .05 level of significance, is the manufacturer's claim correct?

$H_O: P = .90$

$H_{A:} P < .90$

Z tab (α = .05) = -1.64 and reject H_O if Z calc < **-1.64**.

Z calc = (.875 - .90) / √(.90 x .10 /400) = (.875-.90)/.015 = **-1.67**

The corresponding P-value is .0475

Hence, H_O is rejected: the manufacturer's claim is not upheld.

8.7 comparing the means of small samples

We have seen in the preceding sections how the Standard normal distribution can be used to calculate confidence intervals and to carry out tests of significance for the means and proportions of large samples. In this section we shall see how similar methods may be used when we have small samples, using the t-distribution.

The t-distribution

In the previous sections the standard normal distribution (Z-distribution) was used in estimating both point and interval estimates. It was also used to make both one and two-tailed tests. However, it should be noted that the Z-test is applied when the distribution is normal and the population standard deviation σ is known or when the sample size n is large ($\mathbf{n \geq 30}$) and with unknown σ (by taking S as estimator of σ) .

But, what happens when $n<30$ and σ is unknown?

We will use a t-distribution which depends on the number of degrees of freedom (df)..The t-distribution is a theoretical probability distribution (i.e, its total area is 100 percent) and is defined by a mathematical function. The distribution is symmetrical, bell-shaped, and similar to the normal but more spread out. For large sample sizes ($n \geq 30$), both t and Z curves are so close together and it does not much matter which you use. As the degrees of freedom decrease, the t-distribution becomes increasingly spread out compared with the normal. The sample standard deviation is used asan estimate of σ (the standard deviation of the population which is unknown) and appears to be a logical substitute. This substitution, however, necessitates an alteration in the underlying theory, an alteration that is especially important when the sample size, n, is small.

Degrees of Freedom

As explained earlier, the t-distribution involves the degrees of freedom (df). It is defined as the number of values which are free to vary after imposing a certain restriction on your data.

Example: If 3 scores have a mean of 10, how many of the scores can be freely chosen?

Solution

The first and the second scores could be chosen freely (i.e., 8 and 12, 9 and 5, 7 & 15, etc.) But the third score is fixed (i.e., 10, 16, 8, etc.) Hence, there are two degrees of freedom.

Exercise: If 5 scores have a mean of 50, how many of the scores can be freely chosen? Find the degrees of freedom.

Table of t-distributions

The table of t-distribution shows values of t for selected areas under the t curve. Different values of df appear in the first column. The table is adapted for efficient use for either one or two-tailed tests.

Eg1. If df = 8, 5% of t scores are above what value?

Eg2. Find to if n =13 and 95% of t scores are between –to and +to.

Eg3. If df =5, what is the probability that a t score is above 2.02 or below -2.02?

Solutions

1) Look at the table (t-distribution). Along the row labelled "one tail" to the value .05; the intersection of the .05 column and the row with 8 in the df column gives the value of **t = 1.86.**

2. df =13-1 = 12. If 95% of t scores are between -to and + to, then 5% are in the two tails. Look at the table along the row labelled “two tail” to the value .05; the intersection of this .05 column and the row with 12 in the df column gives **to = 2.183**.

3. Two tails are implied. Look along the “df =5” row to find the entry 2.02.

The probability is **.10** .

Computation of Confidence Intervals and Tests of Hypothesis using the t - distribution

Confidence intervals and tests of hypotheses about the mean are carried out with the t distribution just as for the normal distribution, except that we must consider the number of degrees of freedom and use a different table (the table of t distribution).

Eg. The mean pulse rate and standard deviation of a random sample of 9 first year male medical students were 68.7 and 8.67 beats per minute respectively. (Assume normal distribution).

a) Find a 95% C.I. for the population mean.

b) If past experience indicates that the mean pulse rate of first year male medical students is 72 beats per minute, test the hypothesis that the above sample estimate is consistent with the population mean at 5% level of significance.

a) 95% C.I. for the population mean, $\mu = \bar{x} \pm \{t_\alpha \text{ (n-1)df } \times (S/ \sqrt{n})\}$,

where,

t tab (with α = .05 and (n-1)df = $\pm$ 2.31 and $S / \sqrt{9} = 2.89$

Therefore, 95% C.I. for μ = 68.7 $\pm$ (2.31 x 2.89)

= 68.7 $\pm$ 6.7

= (62.0 to 75.4) beats per minute.

b) Hypotheses : H_O : μ = 72

H_A : $\mu \neq 72$

t calc = $(\bar{x} - \mu) / (S/\sqrt{n})$ = $(68.7 - 72) / (8.67/ \sqrt{9})$ = -3.3 / 2 .89 = -1 .14

(This corresponds to a P-value of greater than .10)

t tab (with α = .05 and 8 df) = $\pm$ 2.31

Rule: Reject H_O when t-calculated is either greater than 2.31 or less than –2.31.

(in other words, when the p-value is less than the α - value)

Therefore, the above null hypothesis is accepted.

Conclusion : The result (sample estimate) obtained from the sample is consistent with the population mean. That is, the mean pulse rate of the randomly selected male medical students did not show any significant difference from the total mean pulse rate of male medical students at 5% level of significance.

Paired t-test for difference of means

The characteristic feature of paired samples is that each observation in one sample has one and only one mate or matching observation in the other sample. The comparison of means for paired observations is simple and reduces to the methods already discussed. The key to the analysis is that concern is only with the difference for each pair. The null hypothesis states that the population mean difference ($\mu_d = 0$).

Formula for hypothesis tests**:** $t = (d - \mu_d) / (S_d / \sqrt{n})$ **and C.I. for** $\mu_d = d \pm t_\alpha (S_d / \sqrt{n})$

Where, d = difference in each pair

$\bar{d}$ = the mean of these differences

S_d = the standard deviation of these differences

μ_d = the mean of the population differences

df = n-1 if there are n pairs of scores

Eg. A random sample of 10 young men was taken and the heart rate (HR) of each young man was measured before and after taking a cup of caffeinated coffee. The results were (beats / min):

Subject	HR before	HR after	difference
1	68	74	+6
2	64	68	+4
3	52	60	+8
4	76	72	- 4
5	78	76	- 2
6	62	68	+6
7	66	72	+6
8	76	76	0
9	78	80	+2
10	60	64	+4
Mean	68	71	+3

A) Does caffeinated coffee have any effect on the heart rate of young men ?

(level of significance = .05)

B) Find the 95% C.I. for the mean of the population differences.

solutions

A) $H_O : \mu_d = 0$

$H_{A:}\ \mu_d \neq 0$

$\bar{d} = 3$, $\mu_d = 0$, $S_d = \sqrt{\{\sum(di\text{-}d)^2 / (n-1)\}} = \sqrt{\{ (6-3)^2 + (4-3)^2 + \ldots +(4-3)^2 \} / (10-1)}$

$= \sqrt{(138/9)} = 3.92$

t calc = $(\bar{d} - \mu_d) / (S_d / \sqrt{n}) = (3-0) / (3.92 / \sqrt{10}) = 3 / 1.24 = 2.4$

(This corresponds to a P-value of less than .05)

t tab ($\alpha = .05$, df = 9) = 2.26

t calc is > t tab ⇒ reject H_O.

Hence, caffeinated coffee changes the heart rate of young men.

B) 95% C.I. for the mean of the population differences = d ± 2.26 $(S_d / \sqrt{n})$

$= 3 \pm 2.26(1.24) = 3 \pm 2.8$

$= (0.2, 5.8)$

Exercise Consider the above data on heart rate. Find the confidence intervals and test the hypothesis when the level of significance takes the values .10, .02 and .01. What do you understand from this?

Two means - unpaired t-test (Independent samples)

The unpaired t-test is one of the most commonly used statistical tests. Unless, specifically stated, when a t-test is discussed, it usually refers to an unpaired t-test rather than the paired t-test. A typical research design that uses a t-test is to select a group of subjects and randomly assign them to one of the two groups. Often one group will be a control to whom a placebo drug is given and the other group will receive the drug or treatment to be tested.

Experiments are also conducted in which one group gets a traditional treatment and the other group receives a new treatment to be tested. This is often the case where it is unethical to withhold all treatment. Experiments like this can be single or double blind. Investigators generally have more confidence in double blind tests. These tests should be used whenever it is possible and economically feasible.

Let's repeat the caffeine study but this time we will use an unpaired experiment. We have 20 subjects, all males between the ages 25 and 35 who volunteer for our experiment. One half of the group will be given coffee containing caffeine; the other half will be given decaffeinated coffee as the placebo control. We measure the pulse rate after the subjects drink their coffee. The results are:

Pulse rates in beats / minute

Placebo	**Caffeine**
72	76
76	80
66	78
68	84
68	72
74	66
60	68
64	76
72	76
60	74

Mean	**68**	**75**
Variance	**31.11**	**28.67**

A) Test the hypothesis that caffeine has no effect on the pulse rates of young men ($\alpha = .05$).

B) Find the 95% C.I. for the population mean difference.

Before we perform the unpaired t-test we need to know if we have satisfied the necessary assumptions:

1. The groups must be independent. This is ensured since the subjects were randomly assigned.

2. We must have metric (interval or ratio) data.

3. The theoretical distribution of sample means for each group must be normally distributed (we can rely on the central limit theorem to satisfy this).

4. We need assumption of equal variance in the two groups (Homogeneity of variance).

Since the assumptions are met, we can conduct a two-tailed unpaired t-test .

A) Hypotheses: $H_O : \mu_t = \mu_c$ where, μ_t = population mean of treatment group.

$H_A : \mu_t \neq \mu_c$ μ_c = population mean of control (placebo) group

t calc $=(x_t - x_c) / \sqrt{S^2 (1/n_t + 1/n_c)}$, where, $S^2 = \{(n_t-1)S^2_t + (n_c-1)S^2_c\}/(n_t + n_c - 2)$

S^2 is the pooled (combined) variance of both groups.

n_c = number of subjects in the control group

n_t = number of subjects in the treatment group

x_c = mean of control group

x_t = mean of treatment group

S^2_c = variance of control group

S^2_t = variance of treatment group

$$S^2 = \{ (10-1)\text{x } 28.67 + (10-1) \text{ x } 31.11 \} / 18$$

$$= (258.03 + 279.99)/18 = 538.02 / 18$$

$$= 29.89$$

Therefore, t calc $= (75 - 68) / \sqrt{29.89(1/_{10} + 1/_{10})} = 7 / \sqrt{5.978} = 2.86$

(This corresponds to a P-value of less than .02)

t tab (α = .05 , df = 18) = 2.10

t calc > t tab $\Rightarrow$ reject H_O

Hence, caffeinated coffee has an effect on the pulse rates of young men.

B) 95 % C.I. for the population mean difference = (75-68) ± (2.10 x 2.445)

= 7 ± 5.13 = (1.87, 12.13) beats/ minute. That is, there is a 95% certainty that the population mean difference lies between 1.87 and 12.13 beats / minute.

Exercises

For the data given in the above example,

i) Find the 90% and 99% confidence intervals for the population mean difference.

ii) Test the null hypothesis when α takes the values .1 and .01.

iii) What do you understand from your answers.

8.8 Confidence interval or p – value?

The key question in most statistical comparisons is whether an observed difference between two groups of subjects in a sample is large enough to be evidence of a true difference in the population from which the sample was drawn. As shown repeatedly in the previous sections there are two standard methods of answering this question.

A 95% confidence interval gives a plausible range of values that should contain the true population difference. On average, only 1 in 20 of such confidence intervals should fail to capture the true difference. If the 95% confidence interval includes the point of zero difference then, by convention, any difference in the sample cannot be generalized to the population.

A P-value is the probability of getting the observed difference, or more extreme, in the sample purely by chance from a population where the true difference is zero. If the P-value is greater than 0.05 then, by convention, we conclude that the observed difference could have occurred by chance and there is no statistically significant evidence (at the 5% level of significance) for a difference between the groups in the population.

Confidence intervals and p-values are based upon the same theory and mathematics will lead to the same conclusion about whether a population difference exists. Confidence intervals are preferable because they give information about the size of any difference in the population, and they also (crucially) indicate the amount of uncertainty remaining about the size of the difference.

8.9 Test of significance using the chi-square and fisher's exact tests

8.9.1 The Chi – square test

A chi square (χ^2) distribution is a probability distribution. The chi-square is useful in making statistical inferences about categorical data in which the categories are two and above .

Definition A statistic which measures the discrepancy between K observed frequencies O_1, O_2, . O_k and the corresponding expected frequencies e_1, e_2 . e_k.

$$\textbf{Chi square} = \chi^2 = \sum\{(O_i - e_i)^2\} / e_i$$

The sampling distribution of the chi-square statistic is known as the chi square distribution. As in t distributions, there is a different χ^2 distribution for each different value of degrees of freedom, but all of them share the following characteristics.

Characteristics

1. Every χ^2 distribution extends indefinitely to the right from 0.

2. Every χ^2 distribution has only one (right) tail.

3. As df increases, the χ^2 curves get more bell shaped and approach the normal curve in appearance (but remember that a chi square curve starts at 0, not at $-\infty$)

If the value of χ^2 is zero, then there is a perfect agreement between the observed and the expected frequencies. The greater the discrepancy between the observed and expected frequencies, the larger will be the value of χ^2.

In order to test the significance of the χ^2, the calculated value of χ^2 is compared with the tabulated value for the given df at a certain level of significance.

Example1: In an experiment with peas one observed 360 round and yellow, 130 round and green, 118 wrinkled and yellow and 32 wrinkled and green. According to the Mendelian theory of heredity the numbers should be in the ratio 9:3:3:1. Is there any evidence of difference from the plants at 5% level of significance?

Solution

Hypothesis: **H_O: Ratio is 9:3:3:1**
H_A: Ratio is not 9:3:3:1

Category	O_i	proportion	e_i
RY	360	9/16	360
RG	130	3/16	120
WY	118	3/16	120
WG	32	1/16	40

χ^2 calc = $(360 - 360)^2 / 360 + (130\text{-}120)^2 / 130 + (118\text{-}120)^2 / 120 + (32\text{-}40)^2 /40$

$= 0 + .833 + .033 + 1.60$

$= 2.466 \approx 2.47$

χ^2 tab ($\alpha = .05$, df =3) = 7.8

χ^2 calc < χ^2 tab $\Rightarrow$ accept H_O

Therefore, Ratio is 9:3:3:1.

Example2: The following table shows the relation between the number of accidents in 1 year and the age of the driver in a random sample of 500 drivers between 18 and 50. Test, at a 01 level of significance, the hypothesis that the number of accidents is independent of the driver's age.

There are 75 drivers between 18 and 25 who have no accidents, 115 between 26 and 40 with no accidents, and so on, such a table is called a contingency table. Each “box” containing a frequency is called a cell. This is a 3 x 3 contingency table.

Observed frequencies

Age of driver

Number of accidents	18 - 25	26 - 40	> 40	total
0	75	115	110	300
1	50	65	35	150
≥ 2	25	20	5	50
Total	150	200	150	500

Expected frequencies

Age of driver

Number of accidents	18 - 25	26 - 40	> 40	total
0	90	120	90	300
1	45	60	45	150
≥ 2	15	20	15	50
Total	150	200	150	500

Calculation of expected frequencies: A total of 150 drivers aged 18-25, and 300/500 = 3/5 of all drivers have had no accidents. If there is no relation between driver age and number of accidents, we expect that 3/5(150) = 90 drivers aged 18-25 would have no accidents. I.e.,

$$e11 = \frac{150 \times 300}{500} = 90$$

Similarly,e12(row1 and column 2) = 200x300 /500 = 120

e13 (row1 and column 3) = 150x300 /500 = 90

e22 = (200x150)/500 = 60

e23 = (150x150)/500 = 45

e31 (150x50/500) = 15

e32 = (200x50)/500 = 20

e33 = (150x150)/500 = 15

Hypothesis: **H_O : There is no relation between age of driver and number of accidents**

H_A : The variables are dependent (related)

The degrees of freedom (df) in a contingency table with R rows and C columns is:

$$df = (R - 1)(C - 1)$$

Hence, χ^2 tab with df = 4, at .01 level of significance = 13.3

χ^2 calc = $(75 - 90)^2/90 + (115 - 120)^2/120 + (110 - 90)^2/90 + \ldots + (5 - 15)^2/15$

= 1 + 0.208 + 4.444 + 0.556 + 0.417 + 2.222 + 6.667 + 0 + 6.667

= 22. 2 (This corresponds to a P-value of less than .001)

Therefore, there is a relationship between number of accidents and age of the driver.

8.9.2 **Fisher's exact test**

The chi-square test described earlier is a large sample test. The conventional criterion for the χ^2 test to be valid (proposed by W.G. Cochran and now widely accepted) says that at least 80 percent of the expected frequencies should exceed 5 and all the expected frequencies should exceed 1. Note that this condition applies to the expected frequencies, not the observed frequencies. It is quite acceptable for an observed frequency to be 0, provided the expected frequencies meet the criterion.

If the criterion is not satisfied we can usually combine or delete rows and columns to give bigger expected values. However, this procedure cannot be applied for 2 by 2 tables.

In a comparison of the frequency of observations in a fourfold table, if one or more of the expected values are less than 5, the ordinary χ^2 – test cannot be applied.

The method used in such situations is called Fisher's exact test. The exact probability distribution for the table can only be found when the row and column totals (marginal totals) are given.

Eg1: Suppose we carry out a clinical trial and randomly allocate 6 patients to treatment A and 6 to treatment B .The outcome is as follows:

Treatment type	Survived	Died	Total
A	3	3	6
B	5	1	6
Total	8	4	12

Test the hypothesis that there is no association between treatment and survival at 5% level of significance.

As can be observed from the given data, all expected frequencies are less than 5. Therefore, we use Fisher's exact probability test.

For the general case we can use the following notation:

a	b	r1
c	d	r2
c1	c2	N

The exact probability for any given table is now determined from the following formula:

$$r1!\ r2!\ c1!\ c2!\ /\ N!\ a!\ b!\ c!\ d!$$

The exclamation mark denotes “factorial” and means successive multiplication by cardinal numbers in descending series, that is 5! means 5x4x3x2x1= 120, By convention 0! = 1.

There is no need to enumerate all the possible tables. The probability of the observed or more extreme tables arising by chance can be found from the simple formula given above.

Pr (observed table) = 8! 4! 6! 6! / 12! 3! 5! 3! 1! = .24

Pr (more extreme table) = 8! 4! 6! 6! / 12! 2! 6! 4! 0! = .03

Consequently, the probability that the difference in mortality between the two treatments is due to chance is **2 x (.24 +.03) = .54**

Hence, the hypothesis that there is no association between treatment and survival cannot be rejected.

NB: If the total probability is small (say less than .05) the data are inconsistent with the null hypothesis and we can conclude that there is evidence that an association exists.

8.10 Exercises

1. Consider the following data on living area of mothers and birthweights (kg) of their children which were randomly taken from the records of a given health center.

Mother's living area	Birthweight of child
Rural	2.90 , 3.28, 2.38, 3.06, 2.60, 2.80
Urban	3.19, 3.20, 3.24, 3.16, 2.92, 3.68, 3.40, 3.31, 2.51, 2.80

Test the hypothesis that babies born to mothers coming from rural and urban areas have equal birthweights.

(Assume that the distribution is not skewed and take the level of significance as 5%)

2. Of 30 men employed in a small private company 18 worked in one department and 12 in another department. In one year 5 of the 18 men reported sick with septic hands and of the 12 men 1 did so. What is the

probability that such a difference between sickness rates in the two departments would have arisen by chance?

CHAPTER NINE

CORRELATION AND REGRESSION

9.1 Learning objectives

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

1. Explain the meaning and application of linear correlation
2. Differentiate between the product moment correlation and rank correlation
3. Understand the concept of spurious correlation
4. Explain the meaning and application of linear regression
5. Understand the use of scatter diagrams
6. Understand the methods of least squares

9.2 Introduction

In this chapter we shall see the relationships between different variables and closely related techniques of correlation and linear regression for investigating the linear association between two continuous variables. Correlation measures the closeness of the association, while linear regression gives the equation of the straight line that best describes it and enables the prediction of one variable from the other. For example, in the laboratory, how does an animal's response to a drug change as the dosage of the drug changes? In the clinic, is there a relation between two physiological or biochemical determinations measured in the same patients? In the community, what is the relation between various indices of health and the extent to which health care is available? All these questions concern the relationship between two variables, each measured on the same units of observation, be they animals, patients, or communities. Correlation and regression constitute the statistical techniques for investigating such relationships.

9.3 Correlation Analysis

Correlation is the method of analysis to use when studying the possible association between two continuous variables. If we want to measure the degree of association, this can be done by calculating the correlation coefficient. The standard method (Pearson correlation) leads to a quantity called r which can take any value from -1 to +1. This correlation coefficient r measures the degree of 'straight-line' association between the values of two variables. Thus a value of +1.0 or -1.0 is obtained if all the points in a scatter plot lie on a perfect straight line (see figures).

The correlation between two variables is positive if higher values of one variable are associated with higher values of the other and negative if one variable tends to be lower as the other gets higher. A correlation of around zero indicates that there is no linear relation between the values of the two variables (i.e. they are uncorrelated).

What are we measuring with r? In essence r is a measure of the scatter of the points around an underlying linear trend: the greater thespread of the points the lower the correlation.

The correlation coefficient usually calculated is called Pearson's r or the 'product-moment' correlation coefficient (other coefficients are used for ranked data, etc.).

If we have two variables x and y, the correlation between them denoted by r (x, y) is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} = \frac{\sum xy - [\sum x \sum y]/n}{\sqrt{[\sum x^2 - (\sum x)^2/n][\sum y^2 - (\sum y)^2/n]}}$$

where x_i and y_i are the values of X and Y for the ith individual.

The equation is clearly symmetric as it does not matter which variable is x and which is y (this differs from the case of Regression analysis).

Example: Resting metabolic rate (RMR) is related with body weight.

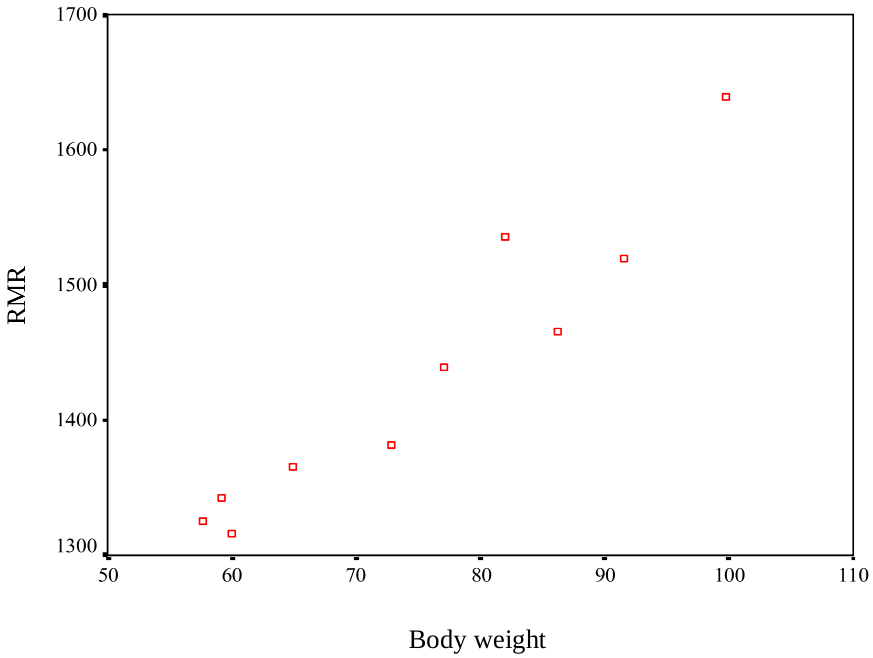
Body Weight (kg)	**RMR (kcal/24 hrs)**
57.6	1325
64.9	1365
59.2	1342
60.0	1316
72.8	1382
77.1	1439
82.0	1536
86.2	1466
91.6	1519
99.8	1639

First we should plot the data using scatter plots. It is conventional to plot the Y- response variable on vertical axis and the independent horizontal axis.

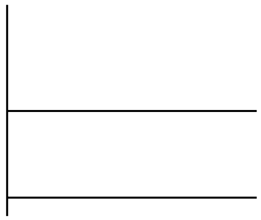
The plot shows that body weight tends to be associated with resting metabolic rate and vice versa. This association is measured by the correlation coefficient, r.

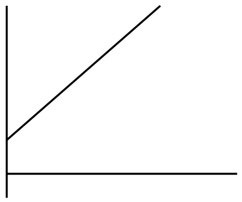
$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \sum (y - \bar{y})^2}}$$

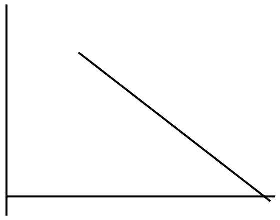
where x denotes body weight and y denotes resting metabolic rate (RMR), and $\bar{x}$ and $\bar{y}$ are the corresponding means. The correlation coefficient is always a number between –1 and +1, and equals zero if the variables are not (linearly) associated. It is positive if x and y tend



to be high or low together, and the larger its value the closer the association. The maximum value of 1 is obtained if the points in the scatter diagram lie exactly on a straight line. Conversely, the correlation coefficient is negative if high values of y tend to go with low values of x, and vice versa. **It is important to note that a correlation between two variables shows that they are associated but does not necessarily imply a 'cause and effect' relationship.**







No correlation (r=0) Imperfect +ve correlation (0<r<1) Imperfect –ve correlation (-1<r<0)

Example: The correlation coefficient for the data on body weight and RMR will be:

$$\sum x = 751.20, \sum x^2 = 58,383.7, \sum y = 14,329, \sum y^2 = 20,634,449, \sum xy = 1,089,9052$$

$$\sum(x-\bar{x})(y-\bar{y}) = \sum xy - (\sum x)(\sum y)/n = 1,089,9052. - (751.2)(1,4329)/10 = 13,510.7$$

$$\sum(x-\bar{x})^2 = \sum x^2 - (\sum x)^2/n = 58,383.7 - (751.2)^2/10 = 1953.56$$

$$\sum(y-\bar{y})^2 = \sum y^2 - (\sum y)^2/n = 20,634,449 - (14,329)^2/10 = 102,424.9$$

$$r = \frac{13,510.72}{\sqrt{(1953.56)1(01,424.9)}} = 0.955$$

Hypothesis test

Under the null hypothesis that there is no association in the population (ρ=0) it can be shown that the quantity $r\sqrt{\frac{n-2}{1-r^2}}$ has a t distribution with n-2 degrees of freedom. Then the null hypothesis can be tested by looking this value up in the table of the t distribution.

For the body weight and RMR data:

$$t = r \times \sqrt{\frac{n-2}{1-r^2}} = 0.955 \times \sqrt{\frac{10-2}{1-(0.955)^2}} = 9.11$$

$P < 0.001$, i.e., the correlation coefficient is highly significantly different from 0.

INTERPRETATION OF CORRELATION

Correlation coefficients lie within the range -1 to +1, with the mid-point of zero indicating no linear association between the two variables. A very small correlation does not necessarily indicate that two variables are not associated, however. To be sure of this we should study a plot of data, because it is possible that the two variables display a non-linear relationship (for example, cyclical or curved). In such casesr will underestimate the association, as it is a measure of linearassociation alone. Consider transforming the data to obtain a linear relation before calculating r.

Very small r values may be statistically significant in moderately large samples, but whether they are clinically relevant must be considered on the merits of each case. One way of looking at the correlation helps to modify over-enthusiasm is to calculate $l00r^2$ (the coefficient of determination), which is the percentage of variability in the data that is 'explained' by the association. So a correlation of 0.7 implies that just about half (49%) of the variability may be put down to the observed association, and so on.

Interpretation of association is often problematic because causation cannot be directly inferred. When looking at variables where there is no background knowledge, inferring a causal link is not justified.

RANK CORRELATION

Occasionally data are not available on the actual measurement of interest, only the relative positions of the members of a group are known. The underlying measurement is continuous - justunobtainable.

For example, James (1985) examined data on dizygotic (DZ) twinning rates and the average daily consumption of milk in 19 European countries in relation to latitude. He was especially interested in twinning rates and latitude.

COUNTRY	Latitude	Twining rate	Average milk consumption
Portugal	40 (1.5)	6.5 (2)	3.8
Greece	40(1.5)	8.8(13)	7.7
Spain	41(3)	5.9(1)	8.2
Bulgaria	42(4)	7.0(3)	
Italy	44(5)	8.6(11.5)	6.5
France	47(6.5)	7.1(4)	10.9
Switzerland	47 (6.5)	8.1 (7.5)	
Austria	48 (8)	7.5 (6)	15.9
Belgium	51 (9.5)	7.3 (5)	11.6
FR Germany	51 (9.5)	8.2 (9)	14.1
Holland	52 (11.5)	8.1 (7.5)	18.9
GDR	52 (11.5)	9.1 (16)	
England/Wales	53 (13.5)	8.9 (14.5)	17.1
Ireland	53 (13.5)	11.0 (18)	24.4
Scotland	56 (15.5)	8.9 (14.5)	
Denmark	56 (15.5)	9.6 (17)	16.8
Sweden	60 (17)	8.6 (11.5)	20.9
Norway	61 (18)	8.3 (10)	19.3
Finland	62 (19)	12.1 (19)	30.4

There are two commonly used methods of calculating rank correlation coefficient, one due to Spearman and one due to Kendall. It is generally easier to calculate Spearman's r_s (also called Spearman's rho). In fact Spearman's rank correlation coefficient is exactly the same as Pearson's correlation coefficient but calculated on the ranks of the observations.

As an alternative approach for hand calculation of Spearman's rank correlation coefficient is given by:

$r_S = 1 - \frac{6\sum d_i^2}{N^3 - N}$, where the difference in ranks, d_i,, is calculated.

This formula bears no obvious similarity to Pearson's but gives identical answers when there are no ties.

The ranks of the data on latitude and DZ twinning rate give thefollowing result:

$$r_S = 1 - \frac{6 \times 366.6}{6859 - 19} = 0.68$$

Although the calculation of r_s should be modified when there are tied ranks in the data, the effect is small unless there are a considerable number of ties. It is probably easier to use the Pearson correlation coefficient on the ranks if relying on a hand calculator or computer. The test of the null hypothesis of no correlation is the same as that for Pearson.

SPURIOUS CORRELATION

The correlation of two variables both of which have been recorded repeatedly over time can be grossly misleading. By such means one may demonstrate relationships between the price of petrol and the divorce rate, consumption of butter and farmers' incomes (a negative relation), and so on. Another example could be the amount of rainfall in Canada and Maize production in Ethiopia (a positive relation).

The same caution applies to studying two variables over time for an individual. Such correlations are often spurious; it is necessary to remove the time trends from such data before correlating them.

9.4 REGRESSION ANALYSIS

The scatter plot of body weight and RMR suggests a linear relationship so we proceed to quantify the relationship between RMR (y) and Body weight (x) by fitting a regression line through the data points. Then the relationship between y and x that is of the following form may be postulated:

$E(Y \mid X) = \alpha + \beta X$ (THAT IS, FOR A GIVEN BODY WEIGHT LEVEL X, THE EXPECTED RESTING METABOLIC RATE $E(Y \mid X)$ IS $\alpha + \beta X$.

Definition: The line **$y = \alpha + \beta x$** is the regression line, where α is the intercept – where the line cuts the y-axis and β is the slope of the line.

The relationship **y =** α **+** β**x** is not expected to hold exactly for every individual, an error term e, which represents the variance of RMR among all individuals with a given body weight level x, is introduced into the model. We will assume that e follows a normal distribution with mean 0 and variance σ^2. The full linear-regression model then takes the following form:

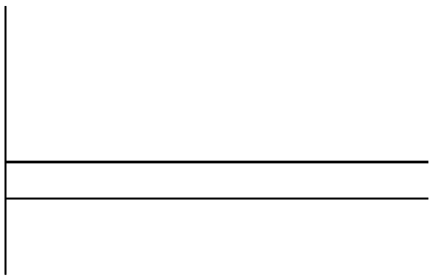
y = α **+** β**x + e**

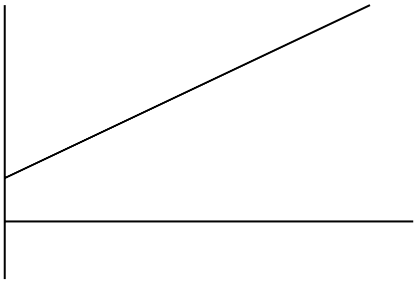
Where e is the residual and it is the part that cannot be accounted for by the model that is normally distributed with mean 0 and variance σ^2.

Definition: For any linear-regression equation of the form **y =** α **+** β**x + e**, y is referred to as the ***dependent*** variable and x as the ***independent*** variable, since we are trying to predict y from x.

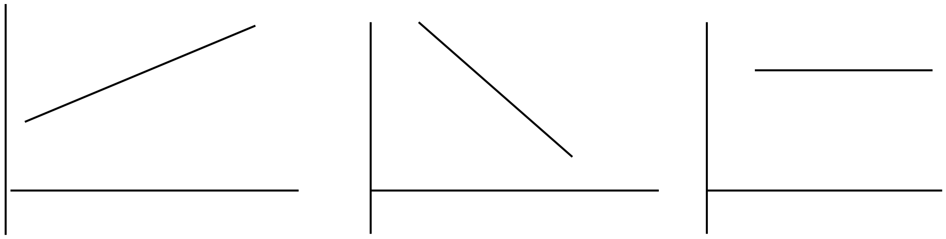
Example: Body weight is the independent variable and resting metabolic rate (RMR) is the dependent variable, since body weight is used to predict the RMR level of individuals.

One interpretation of the regression line is that for a person with body weight of x, the corresponding RMR level will be normally distributed with mean α **+** β**x** and variance σ^2. If σ^2 were 0, then every point would fall exactly on the regression line, whereas the larger σ^2 is the more scatter occurs about the regression line.





How can β be interpreted? If β is greater than 0, then as x increases, the expected value of $y = \alpha + \beta x$ will increase (e.g. relationshipbetween body weight and RMR). If β is less than 0, then as xincreases, the expected value of y will decrease (e.g. relationship between pulse rate and age). If β is equal to 0, then there is no relationship between x and y (e.g. relationship between birth weight and birthday).



(a) $\beta > 0$ (b) $\beta < 0$

(c) $\beta = 0$

FITTING REGRESSION LINE—THE METHOD OF LEAST SQUARES

The question remains as to how to fit a regression line (or, equivalently, to obtain estimates of α and β, denoted by a and b, respectively when the data appear in the form of the previous scatter plot. We would eyeball the data and draw a line that is not too distantfrom any of the points, but this approach is difficult in practice and can be quite imprecise with either a large number of points or a lot of scatter. A better method is to set up a specific criterion that defines the closeness of a line to a set of points and to find the line closest to the sample data according to this criterion.

Consider the data in the above figure and the estimated regression line y = a + bx. The distance d_i of a typical sample point (x_i, y_i) from the line could be measured along a direction parallel to the y-axis. If we let $(x_i, \hat{y}_i)$ = $(x_i, a + bx_i)$ be the point on the estimated regression line at x_i, then this distance is given by $d_i = y_i - \hat{y}_i = y_i - a - bx_i$. A good-fitting line would make these distances as small as possible. Since d_i cannot be 0, the criterion S_1 = sum of the absolute deviations of the sample points from the line $= \sum_{i=1}^{n} |d_i|$ can be used and the line that minimizes S_1 can be found. This strategy has proven to be analytically difficult. Instead, for both theoretical reasons and ease of derivation, the following least-squares criterion is commonly used.

S = sum of the squared distances of the points from the line

$$= \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

Definition: The **least square** line, or **estimated regression line**, is the line y = a + bx that minimizes the sum of squared distances of the sample points from the line given by $S = \sum_{i=1}^{n} d_i^2$. This method of estimating the parameters of the regression line is known as the method of **least squares**.

Based on the least squares estimate, the coefficients of the line y= a + bx are given by:

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x - \bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i) / n}{\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 / n}$$

and by substituting the value of b,

$$a = \bar{y} - b\bar{x}$$

Example: The regression line for the data on body weight and RMR will be

$$\sum x = 751.20, \sum x^2 = 58,383.7, \sum y = 14,329, \sum y^2 = 20,634,44,9, \sum xy = 1,089,9052.$$

$$b = \frac{\sum xy-(\sum x)(\sum y)/n}{\sum x^2 -(\sum x)^2/n} = \frac{10{,}989{,}905.02-(751.2)(14{,}329)/10}{58{,}383.7-(751.2)^2\ /10} = 6.91596$$

$$a = \bar{y} - b\bar{x} = 14{,}329/10 - 6.91596(751.2/10) = 1432.9 - 6.91596(75.12) = 913.3729$$

Thus the regression line is given by $\hat{y} = 913.3729 + 6.91596x$

You recall that we have calculated these results for a random sample of 10 people. Now if we select another sample of 10 people we would get a different estimate for the slope and the intercept. What we are trying to do is to estimate the slope of the "true line". So just like we did when we were estimating the mean or the difference in twomeans etc. we need to provide a confidence interval for the slope of the line.

This is obtained in the usual way by adding to and subtracting from the observed slope a measure of the uncertainty in this value.

In other words, the calculated values for 'a' and 'b' are sample estimates of the values of the intercept and slope from the regression line describing the linear association between x and y in the whole population. They are, therefore, subject to sampling variation and their precision is measured by their standard errors.

$$\text{s.e.}(a) = S \times \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x - \bar{x})^2}} \text{ and s.e.}(b) = \frac{S}{\sqrt{\sum (x - \bar{x})^2}}$$

where

$$S = \sqrt{\frac{\sum (y - \bar{y})^2 - b^2 \sum (x - \bar{x})^2}{n - 2}}$$

S is the standard deviation of the points about the line. It has (n-2) degree of freedom.

Based on the above formulae,

$$\sum (x - \bar{x})^2 = \sum x^2 - (\sum x)^2 / n = 58{,}383.7 - (751.2)^2 / 10 = 1{,}953.556$$

$$\sum (y - \bar{y})^2 = \sum y^2 - (\sum y)^2 / n = 20{,}634{,}449 - (14{,}329)^2 / 10 = 102{,}424.9$$

$$S = \sqrt{\frac{102{,}424.9 - (6.92)^2 (1{,}953.556)}{10 - 2}} = 33.31$$

$$\text{s.e.}(a) = 33.31 \times \sqrt{\frac{1}{10} + \frac{(75.12)^2}{1{,}953.556}} = 57.58$$

$$\text{s.e.}(b) = \frac{33.31}{\sqrt{1{,}953.556}} = 0.754$$

A 95% confidence interval for the long-run slope is, therefore:

Estimated slope ± $t_{1-\alpha/2}$(standard error of slope)

We do not need to know how to calculate the standard error of slope as the computer prints this out for us.

6.92 ± 2.31×(0.754) = 6.92 ±1.76 = (5.18, 8.66)

SIGNIFICANT TEST

Ho: Long-run slope is zero ($\beta=0$)

H_1: Long-run slope is not zero ($\beta\neq0$)

If the null hypothesis is true then the statistic:

$$t = \frac{\text{Observed slope - 0}}{\text{S.E. of obsereved slope}}$$

will follow a t-distribution on n - 2 = 8 degrees of freedom. We lose two degrees of freedom because we have to estimate both the slope and the intercept of the line from the data. A t-distribution on 8degrees of freedom will have 95% of its area between -2.31 and 2.31.A t-test is used to test whether b differs significantly from a specified value, denoted by β.

$$t = \frac{b - \beta}{s.e.(b)}, \ df = n - 2$$

For our data set the calculated t-value is:

$$t = \frac{6.92 - 0}{0.754} = 9.18$$

This is very far out in the right-hand tail and is strong evidence against the hypothesis of no relationship. Notice that the output table gives the t-value and a p-value. (Remember that p is the probability of obtaining our result or more extreme given that the null hypothesis is true (true slope = 0)) Again we would then follow this with a confidence interval for the slope.

PREDICTION

DEFINITION: THE PREDICTED, OR EXPECTED, VALUE OF Y FOR A GIVEN VALUE OF X, AS OBTAINED BY THE REGRESSION LINE, IS DENOTED BY $\hat{y}$=A + BX. THUS THE POINT (X, A + BX) IS ALWAYS ON THE REGRESSION LINE.

In some situations it may be useful to use the regression equation to predict the value of y for a particular value of x, say x'. The predicted value is: y' = a + bx' and its standard error is

$$s.e(y') = S\sqrt{1 + \frac{1}{n} + \frac{(x' - \bar{x})^2}{\sum (x - \bar{x})^2}}$$

Example: What is the expected RMR if a person has body weight of 65 kg?

IF THE BODY WEIGHT WERE 65 KG, THEN THE BEST PREDICTION OF RMR WOULD BE

$\hat{y}$ =913.3729 + 6.91596(65) = 1362.91 kcal/24 hours.

Standard error of prediction is =

$$33.31 \times \sqrt{1 + \frac{1}{10} + \frac{(65 - 75.12)^2}{1{,}953.556}} = 33.31 \times 1.15 = 38.39$$

Assumptions: There are two assumptions underlying the methods of linear regression: Firstly, for any value of x, y is normally distributed and secondly, the magnitude of the scatter of the points about the line is the same throughout the length of the line. This scatter is measured by the standard deviation, S, of the points about the line as defined above. A change of scale may be appropriate if either of these assumptions does not hold, or if the relationship seems non- linear.

PRECAUTIONS IN USE AND INTERPRETATION

1. The relationship must be representable by a straight line. In calculating the correlation coefficient we are presuming that the relationship between the two factors with which we are dealing is one which a straight line adequately describes.
2. The line must not be unduly (extremely) extended. If the straight line is drawn and the regression equation is found, it is dangerous to extend that line beyond the range of the actual observations upon which it is based. For example, in school children height increases with age in such a way that a straight line describes the relationship reasonably well. But to use that line to predict the height of adults would be ridiculous.
3. Association is not necessarily causation. The correlation coefficient is a measure of association, and, in interpreting its meaning, one must not confuse association with causation.

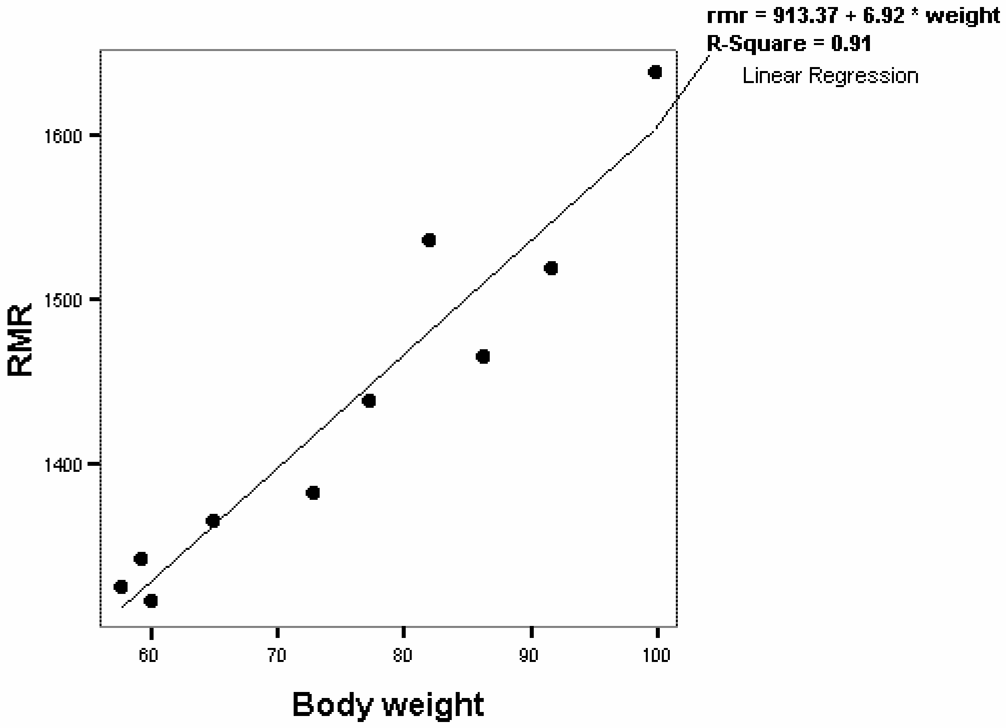
The standard error. As with all statistical values, the correlation coefficient must be regarded from the point of view of sampling errors.

Measure of fit of the model

The R^2 value, of 91.2% provides us with a measure of fit of the model. It is the proportion of variance of Y that can be accounted for by the model.

Assumptions of the Model - Validation of the model

The implication of this model is that all the systematic variation is embodied in the straight line and no further systematic variation should be discernible in the data. If we inspect our scatter plot with the regression line included there are no obvious trends apart fromthe straight line.



It is usual to eliminate the straight-line effect from the data first and then look for systematic patterns in what is left - the residuals. Subtracting from each observed Y value the corresponding value on the line - the fitted Y value does this:

Residual = Y (observed) $-\hat{Y}$ (fitted) = Y $-$ [a + bX].

The following table shows the fitted and observed values for each of the 10 women.

For example Woman 1 has a weight of 57.6 kilos. The predicted/fitted RMR level is

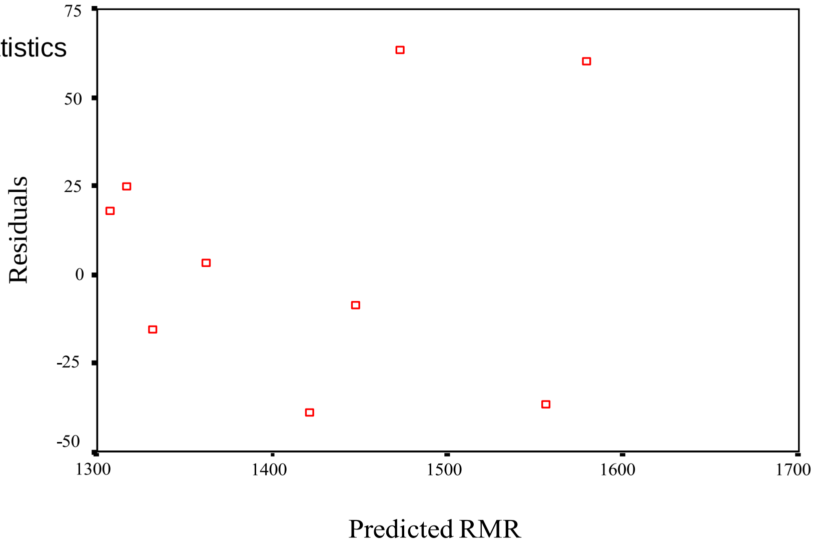
Fitted (predicted) RMR = 913.37 + 6.92*57.6 = 1311.96

Residual = 1325 - 1311.96 = 13.04

Body weight	Observed RMR	Fitted RMR	Residual
57.6	1325	1311.96	13.04
64.9	1365	1362.48	2.52
59.2	1342	1323.03	18.97
60.0	1316	1328.57	-12.57
72.8	1382	1417.15	-35.15
77.1	1439	1446.90	-7.90
82.0	1536	1480.81	55.19
86.2	1466	1509.87	-43.87
91.6	1519	1547.24	-28.24
99.8	1639	1603.99	35.01

Plotting the residuals against anything else should show no systematic patterns. The most useful plot is often that of residuals against the fitted or predicted values.

Scatterplot of residuals by predicted RMR values



The residuals are randomly scattered about zero with no discernible trend with predicted values. They neither increase nor decrease systematically as predicted values increase. Neither is there an indication of any non-linear pattern (there is just a hint of a suggestion that the scatter is greater for larger predicted values than for small ones but with so few observations it would be difficult to draw such a conclusion; we should however bear it in mind if more data are to be collected).

APPENDIX : STATISTICAL TABLES

TABLE 4:AREAS IN ONE TAIL OF THE STANDARD NORMAL CURVE

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.9427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.0	.00130									
3.2	.00069									
3.4	.00034									

3.6	.00016
3.8	.00007
4.0	.00003

Table 5: Percentage points of the t distribution (this table gives the values of t for differing df that cut off specified proportions of the area in one and in two tails of the t distribution)

df	Area in two tails					
	0.2	0.1	0.05	0.02	0.01	0.001
	Area in one tail					
	0.1	0.05	0.025	0.01	0.005	0.0005
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318

13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.358	2.617	3.373
∞	1.280	1.645	1.960	2.326	2.576	3.291

Table 6: Percentage points of the chi-square distribution (this table gives the values of χ^2 for differing df that cut off specified proportions of the upper tail of chi-square the t distribution)

	Area in upper tail					
Df	0.2	0.1	0.05	0.02	0.01	0.001
1	1.642	2.706	3.841	5.412	6.635	10.827
2	3.219	4.605	5.991	7.824	9.210	13.815
3	4.642	6.251	7.815	9.837	11.345	16.268
4	5.989	7.779	9.488	11.668	13.277	18.465
5	7.289	9.236	11.070	13.388	15.086	20.517
6	8.558	10.645	12.592	15.033	16.812	22.457
7	9.803	12.017	14.067	16.622	18.475	24.322

8	11.030	13.362	15.507	18.168	20.090	26.125
9	12.242	14.684	16.919	19.679	21.666	27.877
10	13.442	15.987	18.307	21.161	23.209	29.588
11	14.631	17.275	19.675	22.618	24.725	31.264
12	15.812	18.549	21.026	24.054	26.217	32.909
13	16.985	19.812	22.362	25.472	27.688	34.528
14	18.151	21.064	23.685	26.873	29.141	36.123
15	19.311	22.307	24.996	28.259	30.578	37.697
16	20.465	23.542	26.296	29.633	32.000	39.252
17	21.615	24.769	27.587	30.995	33.409	40.790
18	22.760	25.989	28.869	32.346	34.805	42.312
19	23.900	27.204	30.144	33.687	36.191	43.820
20	25.038	28.412	31.410	35.020	37.566	45.315

21	26.171	29.615	32.671	36.343	38.932	46.797
22	27.301	30.813	33.924	37.659	40.289	48.268
23	28.429	32.007	35.172	38.968	41.638	49.728
24	29.553	33.196	36.415	40.270	42.980	51.179
25	30.675	34.382	37.652	41.566	44.314	52.620

References

1. Colton, T. (1974). Statistics in Medicine, 1st ed. ,Little, Brown and Company(inc), Boston, USA.
2. Bland, M. (2000). An Introduction to Medical Statistics, 3rd ed. University Press, Oxford.
3. Altman, D.G. (1991). Practical Statistics for Medical Research, Chapman and Hall, London.
4. Armitage, P. and Berry, G. (1987). Statistical Methods in Medical Research, 2nd ed. Blackwell, Oxford.
5. Michael, J. (1999). Medical Statistics: A commonsense approach, 3rd ed . Campbell and David Machin.
6. Fletcher, M. (1992). Principles and Practice of Epidemiology, Addis Ababa.
7. Lwanga, S.K. and Cho-Yook T. (1986). Teaching Health Statistics , WHO, Geneva
8. Gupta C.B. (1981). An Introduction To Statistics Methods, 9th Ed. Vikss Publishing House Pvt Ltd, India.
9. Abramson J.H. (1979). Survey Methods In Community Medicine, 2nd Ed. Churchill Livingstone, London and New York.
10. Swinscow T.D.V (1986). Statistics At Square One. Latimer Trend and Company Ltd, Plymouth, Great Britain.
11. Shoukri M.M And Edge V.L (1996). Statistical Methods for Shelath Sciences. CRC Press, London and New York.

12. Kirkwood B.R. (1988). Essentials of Medical Statistics. Blackwell Science Ltd. Australia
13. Spieglman. An Introduction to Demography.
14. Davies A.M And Mansourian (1992). Research Strategies For Health. Publicshed On Behalf of The World Haealth Organization. Hongrefe and Huber Publishers, Lewiston, NY.