FSS 204
Introduction to Statistics in
The Social Sciences
Ibadan Distance Learning Centre Series

FSS 204
Introduction to Statistics in The Social Sciences

By
Arinola, Joshua Ayofe
Department of Economics
University of Ibadan

Published by
Distance Learning Centre
University of Ibadan
# Table of Contents

<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vice Chancellor’s Message…</td>
<td>vi</td>
</tr>
<tr>
<td>Foreword</td>
<td>vii</td>
</tr>
<tr>
<td>General Introduction and Course Objectives</td>
<td>viii</td>
</tr>
<tr>
<td>Lecture One: Introduction to Statistics…</td>
<td>1</td>
</tr>
<tr>
<td>Lecture Two: Data Acquisition process…</td>
<td>18</td>
</tr>
<tr>
<td>Lecture Three: Sampling…</td>
<td>30</td>
</tr>
<tr>
<td>Lecture Four: Classification and Tabulation of Data</td>
<td>55</td>
</tr>
<tr>
<td>Lecture Five: Presentation of Data – Charts and Graphs</td>
<td>69</td>
</tr>
<tr>
<td>Lecture Six: Measures of Central Tendency…</td>
<td>106</td>
</tr>
<tr>
<td>Lecture Seven: Measures of Dispersion…</td>
<td>128</td>
</tr>
<tr>
<td>Lecture Eight: Measures of Relationships…</td>
<td>145</td>
</tr>
<tr>
<td>Lecture Nine: Probability Theory…</td>
<td>164</td>
</tr>
<tr>
<td>Lecture Ten: Probability Distributions…</td>
<td>182</td>
</tr>
<tr>
<td>Lecture Eleven: Estimation…</td>
<td>201</td>
</tr>
<tr>
<td>Lecture Twelve: Tests of Hypothesis…</td>
<td>210</td>
</tr>
<tr>
<td>Lecture Thirteen: The Chi-square distribution…</td>
<td>240</td>
</tr>
<tr>
<td>Lecture Fourteen: Time Series…</td>
<td>257</td>
</tr>
<tr>
<td>Lecture Fifteen: Official Statistics…</td>
<td>281</td>
</tr>
</tbody>
</table>
Vice-Chancellor’s Message

I congratulate you on being part of the historic evolution of our Centre for External Studies into a Distance Learning Centre. The reinvigorated Centre, is building on a solid tradition of nearly twenty years of service to the Nigerian community in providing higher education to those who had hitherto been unable to benefit from it.

Distance Learning requires an environment in which learners themselves actively participate in constructing their own knowledge. They need to be able to access and interpret existing knowledge and in the process, become autonomous learners.

Consequently, our major goal is to provide full multimedia mode of teaching/learning in which you will use not only print but also video, audio and electronic learning materials.

To this end, we have run two intensive workshops to produce a fresh batch of course materials in order to increase substantially the number of texts available to you. The authors made great efforts to include the latest information, knowledge and skills in the different disciplines and ensure that the materials are user-friendly. It is our hope that you will put them to the best use.

Professor Olufemi A. Bamiro, FNSE
Vice-Chancellor
Foreword

The University of Ibadan Distance Learning Programme has a vision of providing lifelong education for Nigerian citizens who for a variety of reasons have opted for the Distance Learning mode. In this way, it aims at democratizing education by ensuring access and equity.

The U.I. experience in Distance Learning dates back to 1988 when the Centre for External Studies was established to cater mainly for upgrading the knowledge and skills of NCE teachers to a Bachelors degree in Education. Since then, it has gathered considerable experience in preparing and producing course materials for its programmes. The recent expansion of the programme to cover Agriculture and the need to review the existing materials have necessitated an accelerated process of course materials production. To this end, one major workshop was held in December 2006 which have resulted in a substantial increase in the number of course materials. The writing of the courses by a team of experts and rigorous peer review have ensured the maintenance of the University’s high standards. The approach is not only to emphasize cognitive knowledge but also skills and humane values which are at the core of education, even in an ICT age.

The materials have had the input of experienced editors and illustrators who have ensured that they are accurate, current and learner friendly. They are specially written with distance learners in mind, since such people can often feel isolated from the community of learners. Adequate supplementary reading materials as well as other information sources are suggested in the course materials.

The Distance Learning Centre also envisages that regular students of tertiary institutions in Nigeria who are faced with a dearth of high quality textbooks will find these books very useful. We are therefore delighted to present these new titles to both our Distance Learning students and the University’s regular students. We are confident that the books will be an invaluable resource to them.

We would like to thank all our authors, reviewers and production staff for the high quality of work.

Best wishes.

Professor Francis O. Egbokhare

Director
General Introduction and Objectives

**Social Statistics** is the use of statistical measurement systems to study human behavior in a social environment. This can be accomplished through polling a particular group of people, evaluating a particular subset of data obtained about a group of people, or by observation and statistical analysis of a set of data that relates to people and their behaviours.

Often, social scientists are employed in the evaluation of the quality of services of a particular group or organization, in analyzing behaviors of groups of people in their environment and special situations, or even in determining the wants or needs of people through statistical sampling.

Statistics and statistical analyses have become a key feature of contemporary social science. Statistics is and has been perhaps most important in economics and psychology that have incorporated and relied primarily on statistical analyses as a method of argument for decades.

Recently, the use of advanced statistical analyses has become popular among the "soft" social sciences such as political science, sociology and anthropology.

There is, however, currently a heated debate regarding the questionable uses and value of statistical methods in social science, especially in political science, with many important statisticians questioning the often broad policy conclusions of political scientists who often misrepresent or misunderstand the limited interpretive power that non-robust statistical methods such as simple and multiple linear regressions allow. Indeed, an important mantra that social scientists cite, but often forget, is that "correlation does not imply causation."

The use of statistics has become so widespread in the social sciences such that many fields like political science, geography, and psychologist have now incorporated advanced causal statistical models that Bayesian methods provide into their syllabi.

**Objectives**

At the end of this course, students are expected to have mastered the following:

1. the aims of statistical techniques in social sciences;
2. basic principles of statistics and its applications in social sciences;
3. the importance of statistics in day to day life of people;
4. define and employ basic statistical terms; and
5. how to conduct a statistical study involving social issues.
LECTURE ONE

Introduction to Statistics

Introduction
The word statistics is derived from the Latin word ‘Statis’ which means a “political state”. Clearly, statistics is closely linked with the administrative affairs of a state such as facts and figures regarding defense force, population, housing, food, financial resources etc. All the attempts made so far in defining statistics point to the same meaning. In this lecture, the true picture of what statistics is all about will be revealed in a clear simple way. Before this, let us highlight what you will come across in the lecture.

Objectives
The readers are expected to have mastered the following after the completion of the lecture:
1. the definition of statistics;
2. the importance of statistics;
3. being able to distinguish the types of statistics; and
4. the types and sources of data.

Pre-Test
1. What is statistics?
2. Mention the types of statistics
3. Enumerates the importance of statistics
4. What are the sources of data?
5. Write short note on: internal and external data.
Define the following: Census, Population, Variable, observation

What are meant by Primary and secondary data?

What are the advantages of primary data?

CONTENT

Nature and Scope of Statistics

The word statistics has several meanings. In the first place, it is a plural noun which describes a collection of numerical data such as population statistics, income and expenditure, employment statistics, accident statistics, imports and exports etc. It is in this sense that the word ‘statistics’ is used by a layman or newspaper.

Secondly, the word statistics as a singular noun is used to describe a branch of applied mathematics whose purpose is to provide methods of dealing with collections of data and extracting information from them in compact form by tabulating, summarizing and analyzing the numerical data or a set of observations.

The word "statistics" is used in several different senses. In the broadest sense, "statistics" refers to a range of techniques and procedures for analyzing data, interpreting data, displaying data, and making decisions based on data. This is what courses in "statistics" generally cover.

In a second usage, a "statistic" is defined as a numerical quantity (such as the mean) calculated in a sample. Such statistics are used to estimate parameters.

The term "statistics" sometimes refers to calculated quantities regardless of whether or not they are from a sample. For example, one might ask about a baseball player's statistics and be referring to his or her batting-average, runs batted in, number of home runs, etc. Or, "government statistics" can refer to any numerical indexes calculated by a governmental agency. Although, the different meaning of “statistics” has the potential for confusion, a careful consideration of the context in which the word is used should make its intended meaning clear. A quantitative measure that describes a characteristic of a sample is called a statistic.
Definitions of Statistics

Statistics has been defined by various authors as:

- Statistics has been defined as the use and development of theory and methods for application in design, analysis and interpretation of information in any substantive field of human endeavours.
- Statistics is the science that deals with collecting and summarizing facts which are expressible in numerical form.
- Statistics is a branch of applied mathematics which utilizes procedures for condensing, describing, analyzing and interpreting set of information.
- It is a way to get information from data.
- Statistics is the study of the methods and theories involved in the collection of data, analysis of data, interpretation of results obtained from the analysis and its utilization.
- Statistics is a tool for creating new understanding from a set of numbers and finally
- Statistics is concerned with scientific method for collecting, organizing, summarizing, presenting and analyzing data, as well as drawing valid conclusions and making reasonable decisions on the basis of such analysis.

The term statistics is used to mean either statistical data or statistical method. When it is used in the sense of statistical data, it refers to quantitative aspect of things and is a numerical description, e.g. the distribution of family incomes or per capital incomes. These are numerical figures. But there are also some that are not but can be made so. For example, sex of babies are not numerical but can be made so. We can count the number of male birth and give the numerical figure of this and likewise the female children.

When statistics is used in terms of statistical method, it is then described as:

1. collection of facts;
2. organization of facts;
3. analysis of facts; and
4. interpretation of facts and on the basis of such interpretation, a conclusion is reached and a reasonable decision is made.

**Types of Statistics**

Statistical investigations and analysis of data fall into two broad categories:

1. Descriptive/deductive statistics and
2. Inferential/inductive statistics.

1. **Descriptive or Deductive Statistics:** These are methods of organizing, summarizing, and presenting data in a convenient and informative way. The major concern of descriptive statistics is to present information in a convenient, useable and understandable form. Descriptive statistics describe the data set that is being analyzed, but does not allow us to draw any conclusions or make any inferences about the data. The techniques of descriptive statistics usually entails the presentation of data in the form of tables and graphs as well as the description of some characteristics of the data by means of numerical values like the averages and measures of variability.

   We can conclude that descriptive statistics is the use of numerical information to summarize, simplify, and present masses of data. It organizes and summarizes data for clearer presentation. This is made for ease of communications. Data on descriptive statistics may come from studies of populations (often called a census study) or samples. The actual method used depends on what information we would like to extract.

2. **Inferential /Inductive Statistics:** This aspect of statistics concerns itself with generalizing the information, or more specifically, with making inferences about population based on data from a sample. This is therefore termed statistical inference. Statistical inference is the process of making an estimate, prediction, or decision about a population based on a sample. We can use statistic to make inferences about parameters. Therefore, we can make an estimate, prediction, or decision about a population based on sample data. Thus, we can apply what we know about a sample to the larger population from which it was drawn.
Importance of Statistics

The importance of statistics can never be overemphasized and the scope of statistics is so vast and ever-increasing. The use of statistics has permeated almost every facet of our lives. It has come to play an important role in almost every field of life and human activity. There is hardly any field where statistical data or methods are used for one purpose or the other. Our arrival in this world and departure from here are recorded as statistical data somewhere and in the same form.

Statistics is an act of making decision in an uncertainty situation. This is the reason why we say that statistics concerns itself with obtaining an insight into the real world by means of the analysis of numerical relationships. Only recently has it been realized that society need not be run on the basis of trial and error. The development of statistics has shown that many aspect of progress depend on the correct analysis of numerical information and relationships particularly in economics, business, and industry. Statistics is a tool of all sciences indispensable to research, intelligent judgments and decision making. In the field of production, statistics play a very important role in guiding the manufacturers on what to produce, how to produce, the quantity and quality of item to produce, when to produce and for whom to produce. Statistical tools are of immense help in quality control, optimisation, inventory level and in dealing with labour problems etc. Production manager looks at quality control data to decide on when to make adjustment in a manufacturing process.

Government, businessmen and individuals collect statistical data required to carry out their activities efficiently and effectively.

Statistics is no doubt an indispensable tool in the planning, building and making of our society. Social scientists have come to the realization of this. Social sciences such as sociology, psychology, political science, economics, geography, anthropology etc. are areas where statistical methods are useful.

Statistics plays an important role in business, because it provides the quantitative basis for arriving at decisions in all matters. It is of high relevance to business in terms of solving problems relating to quality control, market research, inventory control, production planning, auditing, credit management/personnel management, etc. Banks make use of statistics for a number of purposes; statistics has proved to be of immense use in science and social sciences. It has given understanding to the
essential qualities of the laws of nature. In experimental psychology, statistical methods are used in analyzing the experimental data and drawing conclusions there from.

**Purpose of Collecting Data**

Statistics can only deals with numerical data often, however, data which are of qualitative nature can be put into quantitative form e.g., health can be measured by the numbers of days of illness. In business, however, most data can be measured or counted directly, e.g., the number of absentees, sales, output, wages.

Statistics performs three major functions, these are: Descriptive function- which is the way of summarizing information into a useable form; Analytic function- this is achieved by generalization of sample results for the population using the idea of distribution theory and hypothesis testing; Predictive function- This is the main function to predict or forecast.

The following are the roles of statistics:

1. Problem classification and offering necessary solution to it
2. Design of projects
3. Controlling and monitoring of projects
4. Evaluation of the success or failure of programs/projects

From the aforementioned, the purposes of collecting statistics are:

1. To find out facts that are unknown. E.g. hypothesis testing
2. For planning and budgeting
3. For forecasting
4. For controlling
5. For monitoring and evaluation of the on going project.

**Sources of Data**

There are two major sources of statistical data, that is, internal and external sources.

When the data are collected from the organization concerned which can be from both organization’s operating and accounting record, it is
internal data. Most organizations keep such routine data in computer data files for efficient entry, storage and retrieval of such information. These data files together with the associated computer programs constitute the organization’s internal database. If however, data are collected from outside the organization such as journal which is not connected with that organization that collected the data, it is called external data. This external data can as well be divided into two groups- primary and secondary data.

1. **Primary Data**
These are statistical facts, which the investigator originates for the purpose of the inquiry taking place. These are data collected specifically for the purpose in hand. In other words, these are fresh raw facts which the investigator uses his own personal initiative, time, money and materials to obtain. This consists of information collected for an ad-hoc enquiry, that is, for specific purpose. The collection of facts and figures relating to the population in the census provides primary data. The following are the advantages of primary data:

a. Data collected are more complete and exact information needed are obtained.
b. The collected data are reliable and more detailed
c. There are information about the method of collection, presentation and interpretation.
d. Errors which must have been introduced in secondary data such as graphical error may not be present in primary data.

**Disadvantages of Primary Data**
1. It is more expensive to embark on collecting primary data
2. It is time consuming
3. There is possibility of low response

1. **Secondary Data**
These are data extracted from publications containing data previously collected and published by other agency/agencies. It is an administrative source of collecting information which means that the data collected is through administrative by-product. Such data can be obtained from
National Bureau of Statistics (NBS), Central Bank of Nigeria, United Nations or World Bank publications. Let’s quickly go through the advantages and disadvantages.

**Advantages**

1. The information obtained is timely
2. It is less expensive
3. It is faster and easier to collect
4. It gives quick information needed, that is, prompt.
5. It is useful in locating primary source
6. It brings together related data that are scattered in various primary source e.g. United Nations publications, Economic indicator, Annual abstract of statistics, CBN Economic publications and National Bureau of statistics social and Economic Publications.

**Disadvantages**

a. The information is less accurate
b. It is less detailed
c. It is less informative

Before secondary data can be used with safety, these are precaution so as not to obtain useless data.

i. Know the source of data
ii. Know how the information was obtained
iii. The exact definitions of terms should be known
iv. Methods of compilation should be specified.

**Measurement Scales**

Measurement is the assignment of numbers to objects or events in a systematic fashion. Measurement exists in several levels depending on what is to be measured, the instrument to be employed, the degree of accuracy or precision desired and the method of measurement. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio. Measurement in any field of human endeavor involves
the quality or attributes (to which numerical value is assigned in a person or thing).

There is a relationship between the level of measurement and the appropriateness of various statistical procedures. For example, it would be silly to compute the mean of nominal measurements. However, the appropriateness of statistical analyses involving means for ordinal level data has been controversial. One position is that data must be measured on an interval or a ratio scale for the computation of means and other statistics to be valid. Therefore, if data are measured on an ordinal scale, the median but not the mean can serve as a measure of central tendency.

The arguments on both sides of this issue will be examined in the context of an hypothetical experiment designed to determine whether people prefer to work with color or with black and white computer displays. Twenty subjects viewed black and white displays and 20 subjects viewed color displays.

Displays were rated on a 7 point scale where a 1 was the lowest rating and a 7 was the highest rating. This rating scale is only an ordinal scale since there is no assurance that the difference between a rating of 1 and a rating of 2 represents the same degree of difference in preference as the difference between a rating of 5 and a rating of 6.

The mean rating of the color display was 5.5 and the mean rating of the black and white display was 3.9. The first question the experimenter would ask is how likely is it that this big a difference between means could have occurred just because of chance factors such as which subjects saw the black and white display and which subjects saw the color display. Standard methods of statistical inference can answer this question. Assume these methods led to the conclusion that the difference was not due to chance but represented a "real" difference in means. Does the fact that the rating scale was ordinal instead of interval have any implications for the validity of the statistical conclusion that the difference between means was not due to chance?

The answer is an unequivocal "NO." There is really no room for argument here. What can be questioned, however, is whether it is worth knowing that the mean rating of color displays is higher than the mean rating for B & W displays.

The argument that it is not worth knowing assumes that means of ordinal data are meaningless. Supporting the notion that means of ordinal
data are meaningless is the fact that examples can be made up showing that a difference between means on an ordinal scale can be in the opposite direction of what they would have been if the "true" measurement scale had been used.

If means of ordinal data are meaningless, why should anyone care whether the difference between two meaningless quantities (the two means) is due to chance or not? Naturally enough, the answer lies in challenging the proposition that means of ordinal data are meaningless. There are two counter arguments to the example showing that using an ordinal scale can reverse the direction of the difference between means.

The first is philosophical and challenges the validity of the notion that there is some unseen "true" measurement scale that is only being approximated by the rating scale. The second counter argument accepts the notion of an underlying scale but considers the examples to be very contrived and unlikely to occur in real data. Measurement scales used in behavioral research are invariably somewhere between ordinal and interval scales. In the preference experiment, it may not be the case that the difference between the ratings one and two is exactly the same as the difference between five and six, but it is unlikely to be many times larger either. The scale is roughly interval and it is exceedingly unlikely that the means on this scale would favor color displays while the means on the "true" scale would favor the B & W displays.

There are some cases where one can validly argue that the use of an ordinal instead of a ratio scale seriously distorts the conclusions. Consider an experiment designed to determine whether 5-year old children are more distractible than 10-year old children.

<table>
<thead>
<tr>
<th></th>
<th>No Distraction</th>
<th>Distraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>5-year</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>10-year</td>
<td>12</td>
<td>8</td>
</tr>
</tbody>
</table>

It looks as though the 10-year olds are more distractible since distraction cost them 4 points but only cost the 5-year olds 3 points. However, it might be that a change from 3 to 6 represents a larger difference than a change from 8 to 12. Consider that the performance of 5-
year olds dropped 50% from distraction but the performance of 10-year olds dropped only 33%.

Which age group is "really" more distractible? Unfortunately, there is no clearly right or wrong answer. If proportional change is considered, then 5-year olds are more distractible; if the amount of change is considered then 10-year olds are more distractible. Keep in mind that statistical conclusions are not affected by the choice of measurement scale even though the all-important interpretation of these conclusions can be.

In this example, a statistical test could validly rule out chance as an explanation of the finding that 10-year olds lost more points from distraction than did 5-year olds. However, the statistical test will not reveal whether a greater drop necessarily means 10-year olds are more distractible. So, the conclusion that distraction costs 10-year olds more points than it costs 5-year olds is valid. The interpretation depends on measurement issues.

In summary, statistical analyses provide conclusions about the numbers entered into them. Relating these conclusions to the substantive research issues depends on the measurement operations.

Let us consider these scales one by one.

a. **Nominal Scale**

The nominal scale is the simplest and involves only assignment to classes and does not imply magnitude. Examples are the classification of respondents into male and female, teachers classified into trained and untrained. We can thus code the various categories as Male 1, female 0; Trained 1, untrained 0

We record this as nominal data. Note that there is no order of greater than or less than. The nominal scale does not lend itself to some of the useful arithmetical and statistical operations. With nominal scale data, the obvious and intuitive descriptive summary measure is the proportion or percentage of subjects who exhibit the attribute.

b. **Ordinal Scale**

The ordinal scale is based on the order property of real number which says the one real number may be greater than or equal to or
less than another real number. It allows for classification and indication of size of some predefined basis, we can rank people or things according to magnitude. An example is the performance of countries when measuring per capital income, GDP and a host of others. Equal intervals on it do not represent equal quantities, e.g., great, greater greatest.

c. **Interval Scale**

The interval scale possesses the order property of the ordinal scale. The amount of difference between adjacent intervals on the scale is equal. Arithmetic operations permissible on this scale include all those allowed on the ordinal scale; in addition, measurement can be added, subtracted, divided and multiplied by constant yield interpretable results. Comparison between intervals on this scale is meaningful, and is independent of the unit of measurement or the system of assigning scores.

d. **Ratio Scale**

The ratio scale is a scale that has equal intervals as well as new mark which indicate a complete lack of quantity being measured.

**Variable**

Statistical data or information that we gather is obtained by interviewing people, by inspecting items, and in many other ways. The characteristic that is being studied is called a variable. Income of workers in an establishment, ages of politicians, the measurements in geographical surveys, psychologist measurements are examples of variables.

A variable is any measured characteristic or attribute that differs for different subjects. For example, if the weight of 30 subjects were measured, then weight would be a variable.

There are two kinds of variables; these are qualitative and quantitative variables
Quantitative and Qualitative

Variables can be quantitative or qualitative. (Qualitative variables are sometimes called "categorical variables.") Quantitative variables are measured on an ordinal, interval, or ratio scale; qualitative variables are measured on a nominal scale. If five-year old subjects were asked to name their favorite color, then the variable would be qualitative. If the time it took them to respond were measured, then the variable would be quantitative.

A qualitative variable can be identified simply by noting its presence. It describes observations as belonging to one of a set of categories. A quantitative variable consists of values measured on a numerical scale. Data collected for a quantitative variable is often referred to as metric data. For example, the price of a stock (in Naira), the annual income of a family, the volume of sales in a day are examples of quantitative variables.

We can classify quantitative variables further as continuous or discrete. If the variable can assume any numerical value over an interval or intervals, it is a continuous variable. Weight, height and time are examples of continuous variables. The observations can be measured to any degree of accuracy on a numerical scale. Some variables (such as reaction time) are measured on a continuous scale. There is an infinite number of possible values these variables can take on.

In contrast to a continuous variable, a discrete variable is one whose possible values consist of breaks between successive values. Other variables can only take on a limited number of values. For example, if a dependent variable were a subject's rating on a five-point scale where only the values 1, 2, 3, 4, and 5 were allowed, then only five possible values could occur. Such variables are called "discrete" variables. The formal definition is as follows.

A discrete variable is one whose possible values can be counted. It can have a finite number of possible values, or as many values as there are integers. For example, the number of People Democratic Party members in a ward, the income tax paid by a wage earner, production value (in Naira) and sales value which are quantitative variables and are discrete.

Any quantitative measure, that is, a numerical value that describes a characteristic of a population is called a parameter. Parameters are the constants that are peculiar to a given population. A parameter is a numerical quantity measuring some aspect of a population of scores. For
example, the mean is a measure of central tendency. Greek letters are used to designate parameters. At the bottom of this page are shown several parameters of great importance in statistical analyses and the Greek symbol that represents each one. Parameters are rarely known and are usually estimated by statistics computed in samples. To the right of each Greek symbol is the symbol for the associated statistic used to estimate it from a sample.

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Parameter</th>
<th>Statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( \mu )</td>
<td>( m )</td>
</tr>
<tr>
<td>Standard</td>
<td>( \sigma )</td>
<td>( s )</td>
</tr>
<tr>
<td>deviation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion</td>
<td>( \pi )</td>
<td>( p )</td>
</tr>
<tr>
<td>Correlation</td>
<td>( \rho )</td>
<td>( r )</td>
</tr>
</tbody>
</table>

**Independent and Dependent**

When an experiment is conducted, some variables are manipulated by the experimenter and others are measured from the subjects. The former variables are called "independent variables" or "factors" whereas the latter are called "dependent variables" or "dependent measures."

For example, consider a hypothetical experiment on the effect of drinking alcohol on reaction time: Subjects drank water, one beer, three beers, or six beers and then had their reaction times to the onset of a stimulus measured. The independent variable would be the number of beers drunk (0, 1, 3, or 6) and the dependent variable would be reaction time.

The measurement scales can be summarized thus:

**Categorical and Numerical Variables**

1. Any characteristic which varies or changes when moving from individual to individual or object to object in a population.
2. (Qualitative) Variable A variable whose values cannot be interpreted as numbers.

3. (Quantitative) Variable Measurements or counts, values which have meanings as numbers.

4. Set A collection of observations on one or more variables.

Categorical Variable Types
- NOMINAL: Numbers mean nothing other than categorizing, e.g.,
  \[ 0 = \text{Female} \quad 1 = \text{Male} \quad 2 = \text{Divorced/Separated} \]
  \[ 3 = \text{Married} \quad 4 = \text{Widowed} \]
- ORDINAL: Numbers (categories) have a natural ordering, but differences have no meaning, e.g.,
  \[ 1 = \text{Disagree Strongly} \quad 2 = \text{Disagree} \]
  \[ 3 = \text{Agree} \quad 4 = \text{Agree Strongly} \]
  \[ 5 = \text{Poor} \quad 6 = \text{Fair} \]
  \[ 7 = \text{Good} \quad 8 = \text{Excellent} \]

With categorical data, we can replace numbers with letters, symbols, etc.

\[ 0 \ F \ \mathcal{F} \ \text{Female} \]
\[ 1 \ M \ \blacktriangleleft \ \text{Male} \]

NOTE: When analyzing categorical data, we will typically work with counts or percentages of objects or individuals which fall within certain categories.

Numerical Variable Types
1. Interval: The difference between numbers has meaning, e.g., height, weight, temperature.

2. Ratio: Interval data whose ratio also has meaning, e.g., height, weight.

Note: Usually we will group 1 and 2 together and call them "interval-ratio" data. However, there are a few instances where data is interval but not ratio. For example, temperature is interval but not ratio. If the
temperature last night was 40°F and the temperature this afternoon is 80°F, we typically do not say ‘it is 100% hotter this afternoon than it

Continuous and Discrete Data
Data consisting of numerical (quantitative) variables can be further divided into two groups: discrete and continuous.

If the set of all possible values when pictured on the number line, consists only of isolated points, if the set of all values when pictured on the number line, consists of intervals, the most common type of discrete variable we will encounter is a counting variable.

Univariate, Bivariate, and Multivariate Data
Depending on how many variables we are measuring on the individuals or objects in our sample, we will have one of the three following types of data sets:

Measurements made on only one variable per observation.
Measurements made on two variables per observation.
Measurements made on many variables per observation.
In this course, we will concentrate on univariate and bivariate data sets

Summary
In this lecture, you have been able to learn about the roles play by statistics in our day-to-day activities. We have seen statistics as the science that deals with collecting and summarizing facts which are expressible in numerical form. It is a tool of all sciences indispensable to research, intelligent judgments and decision making. You have also seen the extent to which statistical method is required in the analysis and interpretation of figures relating to human behaviours. Two sources of data are identified – primary and secondary sources. Also, there are descriptive and inferential statistics. The functions of statistics such as problem clarification and solutions and a host of beautiful roles being performed by statistics are part of what we have learnt in the lecture.

With all these explanations, the deep understanding of the wonders of statistics have been revealed, the ball is now in your court, kick or abandon. I hope you will surely apply this knowledge to your course of studying.
Post-Test
1. Write short notes on:
   i. Descriptive statistics
   ii. Inferential statistics
   iii. Primary data
   iv. Secondary data
2. Distinguish between:
   a. Continuous and discrete data
   b. Univariate and bivariate data
   c. Interval and ratio scales
   d. Nominal and ordinal scales
   e. Dependent and independent variables.
3. What is a parameter? Differentiate between parameter and statistic.
4. Describe secondary data. Enumerate the advantages and disadvantages of secondary data. What are the precautionary measures we need to take before secondary data can be used?
5. What is a variable? State the purpose of collecting data.

References


Moser, TL: Survey investigation.

LECTURE TWO

Data Acquisition Process

Introduction
In any statistical investigation, the collection of data is the first and the most important matter that must be attended to. Therefore, it is important to give sufficient thought to the selection of appropriate methods of data collection. The decision regarding the choice of the methods should be arrived at after a careful consideration of the nature of information sought, the population under study, accuracy, practicability and cost. The reason for this is that the entire planning and execution of a survey is influenced considerably by methods applied in collecting data. In this lecture, we shall critically examine various methods of acquiring data that reflect closely the relevant characteristics of the population under consideration.

Objectives
At the completion of this lecture, the reader should be able to:
1. plan a survey;
2. collect relevant data for making useful decision(s) at the end;
3. design questionnaire/instrument for your survey; and
4. enumerate various types of survey and their merits.

Pre-Test
1. Define statistical survey
2. a. What are meant by
   i. Population
   ii. Census
   iii. Sample survey
b. What are their merits and demerits?

3. Give five reasons why you think personal interviewing is better than postal questionnaire.

4. What is telephone interviewing? Why do you think telephone interviewing is effective in Nigeria today? Enumerate the merits of this method of collecting data.

5. What do you think should be the qualities of a good interviewer? What are his/her responsibilities?

6. Explain the following:
   a. Direct observation
   b. Participant observation
   What makes the difference between them?

**CONTENT**

**Statistical Survey**

A statistical survey is a process whereby an investigation is carried out in order to solve some social, business, academic and economic problems. It consists of finding facts in particular fields of inquiry. The following are three important types of surveys in which the data collected are of statistical nature: Social surveys, Market research and public opinion polls.

In this nation, there are many organizations that conduct social surveys- Government, Research institutions like NISER, CBN, National Population Commission (NPC), Universities and National Bureau of Statistics (NBS), which is the coordinating body of Nigerian statistical systems. The purpose of this survey is to provide information for other Government departments so that they could carry out their duties more efficiently. This in turn helps federal government to arrive at a policy for solving the particular problems. International bodies like WHO, UNICEF, UNDP, UNIDO, FAO, etc also conduct surveys to solve some problems. Also, some industries and NGOs conduct surveys. Some of the known surveys are sample survey, Census etc. Before we continue, let us discuss some vital issues in conducting survey.
Population
Population is the collection of the individual items, whether of people or things that are to be observed in a given problem situation. It is the collection of all units of a specified type in a region at a particular point or period of time, e.g. human population, houses, farms, patients, financial institutions in the country etc. A population can be finite or infinite. A finite population is a population that has a definite number, e.g., number of pupils in a class, number of banks in the country. An infinite population is a population that is not definite or is uncountable, e.g., number of insects in a farm or number of fish in the sea.

In drawing an inference, it is useful to distinguish between two populations: target population and population sampled. The target population is that population about which an investigator wishes to draw a conclusion e.g. In the investigation to study the adequate remuneration of the academics, the target population here is the academics (the lecturers). The population sampled is that population from which the sample actually was drawn and about which a conclusion can be made e.g. the academics.

Census
A census is the total process of collecting, compiling, evaluating, analyzing and publishing demographic, social and economic data about the entire population of a well defined territory at a specified time.

It is a survey which includes every item or element in the population. It is sometimes called a 100% count. It is the collection of data from the entire statistical population of interest. A complete census is possible if the population is finite.

Advantages of census
The following are the advantages of census:

1. The data can be disaggregated into smaller administration unit e.g. educational aspect, demographic aspect, health aspect, etc.
2. It is efficient for obtaining information of rare events
3. It is useful as a basis for improving current statistics
4. Results obtained are no estimate of the population but the real count or counting.
5. It can be used as a basis for sample survey.
Disadvantages of census

1. It consumes a lot of time, money and labour.
2. Since it requires large number of personnel, supervision will be very difficult.
3. There are possibilities of introducing errors during process

Sample survey

This is an examination of a part of a population to make inference about the population. It is a practical alternative to the complete count, hence in most cases they are used more often than census. One of the goals of a statistical investigation is to explore the characteristics of a large group of items on the basis of a few. Sometimes it is physically, economically, or for some other reason almost impossible to examine each item in a group under study. In such a situation the only recourse is to examine a sub-collection of items from this group. Most scientific research projects, investigations of social problems, and other endeavours lack the resources to carry out complete enumeration of items and can afford to investigate only a small fraction of the population. A small number of units picked from the population for the purpose of acquiring information about the population is called a sample. The following are the merits and demerits of the survey.

Advantages

1. It is economical in terms of time, money and labour. It provides reliable information at much lower cost than census.
2. Results can be obtained more quickly when the data are obtained rapidly and can be analyzed within a short possible time.
3. Higher quality interviewers can be employed since the respondents constitute only a small proportion.
4. It is much easier to follow up non-response by visiting them occasionally.
5. Guiding against incomplete and inaccurate returns is easier.
6. Sample will often provide more accurate information than in census because errors encountered within survey are effectively controlled since the errors can be assessed.
7. Sample survey is the only practical and feasible method to use when investigation entails the destruction of materials.

**Disadvantages**

1. The result is just an estimate which may likely not be the exact result of population.
2. There is a high sampling error.
3. Non-response rate may be high in case the informant is not at home or not willing to respond.

**Methods of Collecting Data**

We have discussed that secondary data is a method of simply abstracting or extracting necessary information from already published statistics. The only method for this is called documentation method.

There are four different methods of collecting primary data, these are: observation method, mail questionnaire method, personal interview and telephone interview.

**a. Observation Method:** Observational method can fairly be called the classic and systematic methods of scientific enquiry. There are two types viz:

i. *Direct observation* and

ii. *Participant observation.*

The third which is sometime used in biomedical survey is the *focus group.*

i. *Direct Observation:* Direct observation implies the use of the eyes rather than the ear and voice. It is defined as accurate watching and noting of phenomena as they occur in nature with regard to causes and effects of mutual relation. The distinguishing feature of observation in the extended sense is that the information required is obtained directly rather than through the report of others. The following are the advantages of the method: It reduces the chance of incorrect data being recorded especially when the informants are unable to provide the information or can give only very inexact answer. This can arise when technical information is needed or when dealing with handicap or in the study of children. It is much more dependable way of
collecting data in the form of measurements of distances and time. It reduces fallibility of memory which can cause reported data to be seriously distorted. It also reduces the unwillingness of informants to give accurate answers to questions. It has the following disadvantages: It is very expensive and uneconomical. It is inappropriate method of studying opinion and attitudes. It could be time consuming. Data on very private matter cannot be obtained via this method. The second one is the participant observation.

ii. **Participant Observation**: In this method the observer joins the daily life of the group or organization he is studying. He watches what happens to the member of the community and how they live and engages in conversation with them to find out the reaction, and to interpret the events that have occurred. The observer’s task is to place himself in the best position for getting complete and unbiased information regarding his study/investigation. The advantages of direct observational method apply to participant observation. There are only two disadvantages which are better referred to as risk effects of participant observation. These are control effect and biased view point effect.

a. **Postal/Mail Questionnaire**: This is a method of data collection whereby questionnaires are sent to respondent by post. The respondents on the other hand fill in the questionnaires unaided by the interviewers and return the completed questionnaires through post to the agency conducting such survey. The postal questionnaire is satisfactory when the law compels the respondents to reply or when sent by association such as Manufacturer Association of Nigeria (MAN) to its members who are usually spread over the entire country. However all sorts of questions can not be asked though mail unless the questions are simple and straightforward. The following are the merits of postal questionnaire:

i. It is cheaper to conduct

ii. It eliminates interviewer bias.

iii. It allows the respondents to remain anonymous.

iv. It gives more time to the respondents to think and give considerable answers to the questions.

v. It is easy to administer
vi. Questionnaire filled by respondent at their own time are more reliable because personal or embarrassing questions would be answered willingly when not scared by an interviewer.

vii. Errors due to reliability and validity of survey results are reduced

viii. The problem of non-contact in a strict sense or respondent not being at home when an interviewer calls is avoided.

A general drawback of this method is that only items of information whose concepts are easily to understand can be included in the questionnaire. Also, it might give rise to a high rate of non-response. Answers to main questionnaire are only accepted as final because there is no opportunity to prove beyond the given answer. Mail questionnaire is inappropriate where spontaneous and immediate answers are wanted. The interviewer cannot be sure that the right person completed the questionnaire. With mail questionnaire, there is no opportunity to supplement the respondent answer by observation. This method is limited to those who can read and write. It may be inefficient where postal system is little developed or underdeveloped.

a. **Telephone Interview**: This method involves asking the necessary questions by telephone. This method has improved tremendously unlike before. The reason is that many people have access to GSM (handset). Telephone interviews are useful for certain kinds of radio or opinion research and this radio research is now easy to conduct because of the advent of the handset. Interview conducted over telephone should usually be kept short and impersonal to maintain the interest of the respondents. The following are the merits of telephone interview: Information collection is faster. It is cheaper than personal interview. It is more flexible than postal questionnaire. Recall to respondent is easier and quicker than by other methods. It is the best method of assessing a difficult area of respondent. There is a higher respondent rate than postal questionnaire. It facilitates recording process without causing any embarrassment to the respondent. It is very suitable for radio and television survey. The demerits are: It is limited to respondents having telephone an obvious evidence of bias. The interviewer can influence the respondent. Cost consideration might limit the number of questions to be asked the respondent and time allotted to answer a given question.
b. **Personal Interview**: This is the method of collecting data especially in social surveys. In it, the interviewer has the major responsibility of locating the individual respondent, interview them and record the supplied answers. The objective of an interview is to elicit frank and complete answer from the respondent. There are two types of interview viz-a-viz formal and informal. There are some necessary conditions for interview to be successful. These conditions are accessibility, cognition and motivation.

   i. **Accessibility**: The interviewer must make sure the required information is accessible to the respondent in order to answer the questions

   ii. **Cognition**: the interviewer as a matter of fact must make the respondent realize or know the role he/she is to play in the interview.

   iii. **Motivation**: there must be motivation on the part of the respondent to answer questions accurately. The initial motivation in many interviews is not more than simple courtesy.

**Qualities of the Interviewers**

There are some qualities an interviewer must possess in order to perform effectively. These qualities are:

   a. **Honesty**: He must be honest, scrupulous and be a person of high integrity.

   b. **Accuracy**: Interviewer should be accurate in their recording of answer in the way they follow instructions, apply definitions and carried the administrative duties.

   c. **Adaptability**: the interviewer should be able to adapt himself to any varying circumstances.

   d. **Personality and Temperance**: the interviewer’s personality should neither be overaggressive nor over sociable. Pleasantness and business-like manner is the ideal combination.

   e. **Intelligence and Education**: The interviewer must have sufficient intelligent to understand and follow complicated instructions and to be adaptable within given limits.
Advantages of Personal Interview

The following are the primary advantages of personal interviewer

1. The response rate is usually improved when people are confronted in person.
2. It allows more accurate information to be obtained
3. It can be used for person of all educational levels.
4. It can be used to explore areas in which little information exists.
5. Question misinterpretation is eliminated by interviewer’s assistance, thus providing some level of uniformity in supplied answers.
6. A skilled interviewer can easily persuade unwilling respondent thereby increase the number of responses.
7. Interviewer can know the reaction of respondent to question if necessary.

The major limitations of this method include among other things

1. High cost of survey.
2. Errors in recording the response and interviewer/respondent biases.
3. It may be difficult to interview some individual such as highly placed company executives, people in high income group. Respondents may give inaccurate or false information due to lapse in memory, misunderstanding the question or it may be deliberate.

Questionnaire Design Techniques

An important step in any data collection program is the design and preparation of questionnaires. Questionnaires are designed to communicate requests to respondents for information and cues to help respondents produce accurate responses, and to motivate them to work diligently to provide the information requested. Designing a good questionnaire involves selecting the questions needed to meet the research objectives, testing them to make sure they can be asked and answered as planned, then putting them into a form to maximize the ease with which respondents and interviewers can do their jobs. It is a basic tool to the enumerator.
Basic Steps

The major objectives of questionnaire design are to facilitate interviewing and response to collect the required information and facilitate processing. But before these can be achieved, we need to understand the basic steps in design a questionnaire. These steps are:

1. Type of data required: This refers to the population of interest or target population.
2. Type of information required should be specified e.g. whether educational level, income, marital status etc.
3. Write up the first draft of your questionnaire. In doing this, in a big survey, personnel in various fields must be there in order to have questionnaire that will satisfy the objectives mentioned above.
4. Re-examine and revise the draft
5. Pre-test the draft to see how well the questions are understood and interpreted and respondent’s attitudes to question.
6. Improvement (re-editing) of draft so as to produce a final draft.

Guidelines for Questionnaire Design

The following are the guidelines for questionnaire design:

1. **Decision on forms**: The form/questionnaire should be attractive, especially if respondents have to answer unaided. Preferably short questions which should be cleared and simple.
2. **Decision on the number of questions**: This naturally depends on the number of topics or variables of interest. Ideally, questions should be few as possible. Too many and lengthy questions may discourage the respondents.
3. **Decision on question order**: Question should be arranged logically, one question leading to the other. Specific question should always follow general one. The opening question should be very interesting. This will ensure that the respondent cooperates in the survey. Ask very intimate question at the very end.
4. **Decision on the question contents**: Only necessary questions should be asked. Questions should be within the experience and competence of the respondents and questions should not be loaded in one direction.
5. **Decision on question wording**: The question wording should be simple and framed in a way to be interpreted in the same manner by all. Avoid ambiguous statements, terminologies or antiquated languages.

6. **Decision on question type**: Decide on the type of question type, that is, closed or fixed response type which is structured with response categories or unstructured called open-ended response question. However, in questionnaire design closed ended or fixed response is preferred because of the following:
   - It facilitates coding
   - Interviewing is faster
   - It saves space
   - Response is fast
   - Provides uniform or standardized answers
   - Reduces the tendency to error.

**Summary**
So far, we have seen how to collect data. The design of instrument to be used has been clearly spelt out and the technique involved. The methods as explained are five in number, documentary method for collecting secondary data and direct observation, personal interview, telephone interview, and mail questionnaire methods for primary data.

**Post-Test**
1. The following are various types of survey
   i. Sample survey
   ii. Census
   Define each of them and enumerate their merits and demerits
2. Define questionnaire and what are the general guidelines for designing a good questionnaire?
3. Why is closed-ended or fixed response questionnaire is preferable to open-ended?
4. State the objectives of questionnaire design. Enumerate the basic steps in questionnaire design technique.
5. Distinguish between the mail questionnaire and telephone interview.

References


LECTURE THREE

Sampling

Introduction
Instead of obtaining data from the whole of the material being investigated, sampling methods are often used in which only a sample selected from the whole is dealt with, and from this sample conclusions are drawn relating to the whole. If the conclusions are to be valid, the sample must be representative of the whole. The selection of this sample must therefore be made with great care.

Objectives
At the end of this lecture, you should be able to:
1. discuss the meaning of sampling;
2. discuss the reasons and basis for sampling;
3. discuss the sampling errors, sample design and elements in the design;
4. discuss bias in sampling; and
5. discuss types of sampling

Pre-Test
1. What is a statistical sampling? Discuss the essentials of the various methods of sampling, state and explain the two laws upon which the sampling technique depends.
2. Why do we sample and what are the bases of sampling?
3. Explain the use of sampling in statistical investigations. What considerations should principally be borne in mind when making a social inquiry by sample?
Samples and Sampling

Sampling is used to make inferences about a population from a relatively small number of observations that are assumed to be representative of the population.

“Statistical designs always involve compromises between the desirable and the possible.”

As the quote above from Leslie Kish highlights, all research designs involve some form of compromise or adjustment. One of the dimensions on which such compromises are made relates to the populations about which we wish to learn. There are many research questions we would like to answer that involve populations that are too large to consider learning about every member of the population. How have wages of Nigerians workers changed over the past ten years? How do Nigerians feel about the job that the President is doing? What are the management practices of banking industries?

Questions such as these are important in understanding the world around us, yet it would be impractical, if not impossible, to measure the wages of all Nigerian workers, the feelings about the President of Nigeria, and the banking practices of the world's banks. Generally, in answering such questions, social scientists examine a fraction of the possible population of interest, drawing statistical inferences from this fraction. The selection process used to draw such a fraction is known as sampling, while the group contained in the fraction is known as the sample.

It is not only statisticians or quantitative researchers that sample. Journalists who select a particular case or particular group of people to highlight in a news story are engaging in a form of sampling. Most of us, in our everyday lives, do some sampling, whether we realize it or not. Although, you may not have listened to all the songs of a particular band or singer, you likely would be able to form an opinion about such songs from hearing a few of them. In making such inferences you've relied on a subset of entities (some songs of an artist) to generalize to a larger group (all songs by an artist). You've sampled.
Why Sample?

In a social researcher's ideal world all data would be collected by census. However, in practice this approach is not always practical because of time and cost constraints. As an alternative sampling methods are used to provide estimates based on data from a very small percentage of the population.

Sampling is done in a wide variety of research settings. Listed below are a few of the benefits of sampling:

1. **Reduced cost**: It is obviously less costly to obtain data for a selected subset of a population, rather than the entire population. Furthermore, data collected through a carefully selected sample are highly accurate measures of the larger population. Public opinion researchers can usually draw accurate inferences for the entire population of Nigeria from interviews of only 1,000 people.

2. **Speed**: Observations are easier to collect and summarize with a sample than with a complete count. This consideration may be vital if the speed of the analysis is important, such as through exit polls in elections.

3. **Greater scope**: Sometimes highly trained personnel or specialized equipment limited in availability must be used to obtain the data. A complete census (enumeration) is not practical or possible. Thus, surveys that rely on sampling have greater flexibility regarding the type of information that can be obtained.

4. **Reliability**: A high level of reliability can be achieved because fewer units are surveyed in a sample than in a full survey and therefore resources can be concentrated on obtaining reliable information.

5. **Resource allocation**: by using samples it is possible to carry out several studies concurrently, and therefore use resources efficiently.

It is important to keep in mind that the primary point of sampling is to create a small group from a population that is as similar to the larger population as possible. In essence, we want to have a little group that is like the big group. With that in mind, one of the features we look for in a sample is the degree of representativeness - how well does the sample
represent the larger population from which it was drawn? How closely do the features of the sample resemble those of the larger population?

There are, of course, good and bad samples, and different sampling methods have different strengths and weaknesses. Before turning to specific methods, a few specialized terms used in sampling should be defined.

**How does Sampling Work?**

Sampling works because, it is not necessary to collect data from all people in order to generate statistics about that population.

After a certain sample size, there is no need to collect more data. The extra data does not improve the accuracy of the estimate to any great extent.

**Sampling Terminology**

Samples are always drawn from a population, but we have not defined the term "population." By "population" we denote the aggregate from which the sample is drawn. The population to be sampled (the sampled population) should coincide with the population about which information is wanted (the target population). Sometimes, for reasons of practicality or convenience, the sampled population is more restricted than the target population. In such cases, precautions must be taken to secure that the conclusions only refer to the sampled population.

Before selecting the sample, the population must be divided into parts that are called sampling units or units. These units must cover the whole of the population and they must not overlap, in the sense that every element in the population belongs to one and only one unit. Sometimes the choice of the unit is obvious, as in the case of the population of Americans so often used for opinion polling. In sampling individuals in a town, the unit might be an individual person, the members of a family, or all persons living in the same city block. In sampling an agricultural crop, the unit might be a field, a farm, or an area of land whose shape and dimensions are at our disposal. The construction of this list of sampling units, called a frame, is often one of the major practical problems.
The Gallup Poll a Familiar Example of Sampling
Most of us are familiar with sampling at some level through seeing reports about levels of popular opinion about some current topic. Newspapers, television and nowadays GSM (phone call) programs are filled with references to the current state of popular opinion.

One of the most prestigious firms in the polling business is The Gallup Organization. Their web site has an excellent description of why sampling is common in social science research, how it is conducted, and some other issues. For example, the page contains the following statement about the value of sampling.

The basic principle: a randomly selected, small percent of a population of people can represent the attitudes, opinions, or projected behaviour of all of the people, if the sample is selected correctly.

The Basis of Sampling
The possibility of reaching valid conclusions concerning a population from a sample is based on two general laws:

a. The law of statistical regularity: this states that a reasonably large sample selected at random from a large population will be on average, representative of the characteristics of the population.

b. The second one is the law of the inertia of large numbers: this states that large groups of data show a higher degree of stability than small ones. There is tendency for variations in the data to be cancelled out by each other.

Sampling Errors
Sampling error is an error attributable to sampling. It is a general term. The errors vary depending on the size of the sample, sampling technique or procedure involve, the extent to which the materials vary, method of data collection, compilation and computation and choice of sampling frame.

Non-sampling Errors
These are due to problems involved with the sample design. Many of them would arise with a full survey, but some of them are due specifically to
Sample design, including such factors as the choice of a sampling frame and units. Non-sampling errors arise due to deliberate selection, substitution, failure to cover whole sample, etc.

Sample Design

a. Principles of sample design
   i. To avoid bias in the selection procedure
   ii. To achieve the maximum precision for a given outlay of money and time.

b. Planning and Execution of sample survey
   The following are procedural steps in planning and execution of a survey:
   1. decide on the objective of the survey;
   2. know the population to be covered;
   3. select the right sampling frame;
   4. select the sampling unit;
   5. decide on sampling selection;
   6. choose a sampling method;
   7. information to be collected;
   8. method of collecting your data;
   9. decide on time reference and reference period;
   10. decide on questionnaire/schedule design;
   11. training of interviewers and their supervisors;
   12. pilot study/survey;
   13. field work and inspection of returns;
   14. solving the problem of non-response; and
   15. analysis of data.

c. Elements in the sample design
   i. The sampling population is the group of people, items or units under investigation.
   ii. The sample units are the people or items which are to be sampled.
iii. The sampling frame is the list of people or items or units from which the sample is taken or it is a list of the names of every item in the population. It should be comprehensive, complete and up-to-date to keep bias to a minimum. Examples of sampling frames include electoral register, telephone directories, wage lists, etc.

iv. The survey method includes designing questionnaires and deciding how to distribute them or how to carry out interviews or observations.

v. Sampling methods fall into two categories: random sampling and non-random sampling (these are fully discussed below).

**Sample Size**

The main decision needed in deciding on a sample design is sample size. To decide on this, a number of questions need to be decided on:

a. What are the key estimates for the study?

b. How precise do those estimates need to be? (i.e. what size of standard error or confidence interval can be tolerated?)

c. Are there key sub-groups for which separate estimates will be needed?

d. Does the survey need to be large enough to detect change over time between surveys, or differences between key sub-groups?

**Bias in Sampling**

Population estimates based on survey data will be inaccurate if the sample is biased. There are a number of reasons why a sample may be biased:

- Sample bias occurs when the selected sample is **systematically different** to the population. The sample must be a fair representation of the population we are interested in.
- The sample size may be **too small** to produce a reliable estimate.

There may be **variability in the population**. If you want your sample to give an estimate that is close to the population value, you need to take into account how much variability there is in the variable you are trying to
measure. All else being equal, the greater the variability the larger the sample size needed.

Bias may arise from:

1. **The sampling frame**: if it does not cover the population adequately and accurately.
2. **Non-response**: if some sections of the population are impossible to find or refuse to cooperate.
3. **The sample**: if the most ‘convenient’ sample is selected it may be biased and non-random.
4. **Question wording**: poorly worded, ambiguous questions and interviewer bias may cause problem in sample surveys.
5. **The sampling unit**: a personal element may enter into selection. Substituting one unit or person for another may introduce bias.

Types of Sampling

We may then consider different types of probability samples. Although there are a number of different methods that might be used to create a sample, they generally can be grouped into one of two categories: **probability samples** or **non-probability** samples.

**Probability Samples**

The idea behind this type is random selection. More specifically, each sample from the population of interest has a known probability of selection under a given sampling scheme. There are five main types of probability sample. The choice of these depends on nature of research problem, the availability of a good sampling frame, money, time, desired level of accuracy in the sample and data collection methods. Each has its advantages, each its disadvantages. They are as described below.

**Simple Random Sampling**

The most widely known type of a random sample is the simple random sample (SRS). This is characterized by the fact that the probability of selection is the same for every case in the population. Simple random sampling is a method of selecting n units from a population of size N such that every possible sample of size n has equal chance of being drawn.
Characteristics:
1. Each person has same chance as any other of being selected
2. Standard against which other methods are sometimes evaluated
3. Suitable where population is relatively small and where sampling frame is complete and up-to-date

Procedure:
1. Obtain a complete sampling frame
2. Give each case a unique number, starting at one
3. Decide on the required sample size
4. Select that many numbers from a table of random numbers or using computer

Table of random numbers (usually found at back of statistics textbooks) e.g.

```
92941 04999 77422 25992 27372
94157 43252 83266 47196 94045
48135 34237 46293 46178 50110
78907 37586 50940 88094 28209
82843 43383 32561 62108 46076
```

Decide on a pattern of movement through table and stick to it, e.g. numbers from every second column and every row. If a number comes up twice or a number is selected which is larger than population number, discard it.

Example 2
Example of simple random sampling of 10 households from a list of 40 households.

We have a list of 40 heads of households. Each has a unique number, 1 through 40. We want to select 10 households randomly from this list. Using a random number table, we select consecutive 2-digit numbers starting from the upper left. If a random number matches a household's number, that household is added to the list of selected households. If a
random number does not match a household's number (for example, if it is greater than 40), then it does not select a household. After each random number is used, it is crossed out so that it is never used again. We continue to select households until we have 10.

<table>
<thead>
<tr>
<th>Heads of households</th>
<th>Random number table</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Imaani M</td>
<td>9947 2885 5850 1927 2254 8884 6086 4310</td>
</tr>
<tr>
<td>2 Tsabi A</td>
<td>2426 4764 7274 4600 4735 9566 6435 3152</td>
</tr>
<tr>
<td>3 Mereu G</td>
<td>7000 2865 8357 4475 9604 0042 1108 7949</td>
</tr>
<tr>
<td>4 Aaron L</td>
<td>2932 8958 9582 2235 1140 1164 7841 1688</td>
</tr>
<tr>
<td>5 Obana V</td>
<td>4097 8965 5030 1785 5420 0125 4953 1332</td>
</tr>
<tr>
<td>6 Apen E</td>
<td>5540 6278 1584 4982 3258 1374 1617 7427</td>
</tr>
<tr>
<td>7 Mobong P</td>
<td>3320 8788 7958 9615 9862 7960 8140 8607</td>
</tr>
<tr>
<td>8 Ngara T</td>
<td>8077 2085 2580 2091 8921 0970 3134 8441</td>
</tr>
</tbody>
</table>

Note that even though the selected households appear somewhat clustered, if the random number table is truly random, the selected households have been randomly selected.

Another example may make this easier to understand. Imagine you want to carry out a survey of 100 voters in a small town with a population of 1,000 eligible voters. With a town this size, there are "old-fashioned" ways to draw a sample. For example, we could write the names of all voters on a piece of paper, put all pieces of paper into a box and draw 100 tickets at random. You shake the box, draw a piece of paper and set it aside, shake again, draw another, set it aside, etc. until we had 100 slips of
paper. These 100 form our sample. And this sample would be drawn through a simple random sampling procedure - at each draw, every name in the box had the same probability of being chosen.

In real-world social research, designs that employ simple random sampling are difficult to come by. We can imagine some situations where it might be possible - you want to interview a sample of doctors in a hospital about work conditions. So you get a list of all the physicians that work in the hospital, write their names on a piece of paper, put those pieces of paper in the box, shake and draw. But in most real-world instances it is impossible to list everything on a piece of paper and put it in a box, then randomly draw numbers until desired sample size is reached.

There are many reasons why one would choose a different type of probability sample in practice.

Example 1
Let's suppose your sampling frame is a large city's telephone book that has 2,000,000 entries. To take a SRS, you need to associate each entry with a number and choose n=200 numbers from N=2,000,000. This could be quite an ordeal. Instead, you decide to take a random start between 1 and N/n=20,000 and then take every 20,000th name, etc. This is an example of systematic sampling, a technique discussed more fully below.

Example 2
Suppose you wanted to study dance club and bar employees in NYC with a sample of n = 600. Yet there is no list of these employees from which to draw a simple random sample. Suppose you obtained a list of all bars/clubs in NYC. One way to get this would be to randomly sample 300 bars and then randomly sample 2 employees within each bars/club. This is an example of cluster sampling. Here the unit of analysis (employee) is different from the primary sampling unit (the bar/club).

In each of these three examples, a probability sample is drawn, yet none is an example of simple random sampling. Each of these methods is described in greater detail below.

Although simple random sampling is the ideal for social science and most of the statistics used are based on assumptions of SRS, in practice, SRS are rarely seen. It can be terribly inefficient, and particularly difficult
when large samples are needed. Other probability methods are more common. Yet SRS is essential, both as a method and as an easy-to-understand method of selecting a sample.

To recap, though, that simple random sampling is a sampling procedure in which every element of the population has the same chance of being selected and every element in the sample is selected by chance.

**Stratified Random Sampling**

In this form of sampling, the population is first divided into two or more mutually exclusive segments based on some categories of variables of interest in the research. It is designed to organize the population into homogenous subsets before sampling, then drawing a random sample within each subset. With stratified random sampling the population of $N$ units is divided into subpopulations of units respectively. These subpopulations, called *strata*, are non-overlapping and together they comprise the whole of the population. When these have been determined, a sample is drawn from each, with a separate draw for each of the different strata. The sample sizes within the strata are denoted by respectively. If a SRS is taken within each stratum, then the whole sampling procedure is described as stratified random sampling.

The primary benefit of this method is to ensure that cases from smaller strata of the population are included in sufficient numbers to allow comparison. An example makes it easier to understand. Say that you’re interested in how job satisfaction varies by race among a group of employees at a firm. To explore this issue, we need to create a sample of the employees of the firm. However, the employee population at this particular firm is predominantly white, as the following chart illustrates:

![Racial Distribution of Firm](chart.png)
If we were to take a simple random sample of employees, there's a good chance that we would end up with very small numbers of Blacks, Asians, and Latinos. That could be disastrous for our research, since we might end up with too few cases for comparison in one or more of the smaller groups.

Rather than taking a simple random sample from the firm’s population at large, in a stratified sampling design, we ensure that appropriate numbers of elements are drawn from each racial group in proportion to the percentage of the population as a whole. Say we want a sample of 1000 employees - we would stratify the sample by race (group of White employees, group of African American employees, etc.), then randomly draw out 750 employees from the white group, 90 from the African American, 100 from the Asian, and 60 from the Latino. This yields a sample that is proportionately representative of the firm as a whole.

Stratification is a common technique. There are many reasons for this, such as:

- If data of known precision are wanted for certain subpopulations, than each of these should be treated as a population in its own right.
- Administrative convenience may dictate the use of stratification, for example, if an agency administering a survey may have regional offices, which can supervise the survey for a part of the population.
- Sampling problems may be inherent with certain subpopulations, such as people living in institutions (e.g. hotels, hospitals, prisons).
- Stratification may improve the estimates of characteristics of the whole population. It may be possible to divide a heterogeneous population into subpopulations, each of which is internally homogenous. If these strata are homogenous, i.e., the measurements vary little from one unit to another; a precise estimate of any stratum mean can be obtained from a small sample in that stratum. The estimate can then be combined into a precise estimate for the whole population.
- There is also a statistical advantage in the method, as a stratified random sample nearly always results in a smaller variance for the estimated mean or other population parameters of interest.

**Stratification**

Stratification essentially means dividing the sampling frame into groups (strata) before sampling. Stratification reduces the risk of drawing an extreme sample, unrepresentative of the population.
A simple example would be to take a sampling frame of, say, business establishments and then to sort them into size strata before sampling. The sample would then be described as a sample stratified by size.

There are two methods of stratified sampling: proportionate and disproportionate.

In a **proportionate** stratified sample the sampling frame is divided into strata but the same sampling fraction is applied per stratum. This means that each stratum is sampled from in its correct proportion.

In a **disproportionate** stratified sample the sampling fraction differs between strata. This means that individuals from the strata with the highest sampling fractions will be over-represented in the sample.

Disproportionate sampling is generally used when there is a need to boost the sample size within a particular stratum or strata (e.g. for boosting the number of young people or minority ethnic groups).

**Systematic Sampling**

This method of sampling is at first glance very different from SRS. In practice, it is a variant of simple random sampling that involves some listing of elements - every nth element of list is then drawn for inclusion in the sample. Say you have a list of 10,000 people and you want a sample of 1,000.

Creating such a sample includes three steps:

- Divide number of cases in the population by the desired sample size. In this example, dividing 10,000 by 1,000 gives a value of 10.
- Select a random number between one and the value attained in Step 1. In this example, we choose a number between 1 and 10 - say we pick 7.
- Starting with case number chosen in Step 2, take every tenth record (7, 17, 27, etc.).

More generally, suppose that the N units in the population are ranked 1 to N in some order (e.g., alphabetic). To select a sample of n units, we take a unit at random, from the 1st k units and take every k-th unit thereafter.
Steps in selecting a systematic random sample:

Calculate the sampling interval (the number of households in the population divided by the number of households needed for the sample)

Select a random start between 1 and sampling interval

Repeatedly add sampling interval to select subsequent households.

Example of systematic random sampling of 10 households from a list of 40 households.

We first calculate the sampling interval by dividing the total number of households in the population (40) by the number we want in the sample (10). In this case, the sampling is 4. We then select a number between 1 and the sampling interval from the random number table (in this case 3). Household #3 is the first household. Then count down the list starting with household #3 and select each 4th household. For example, the second selected household is 3 + 4, or #7. Note that when you reach the end of the list, you should have selected your desired number of households. If you have not, you have counted wrong or miscalculated the sampling interval. You should go back and start over.

This is what your final selection should look like:

<table>
<thead>
<tr>
<th>Households</th>
<th>Random number table</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Isiere M</td>
<td>21 Atem E</td>
</tr>
<tr>
<td>2 Tabere C</td>
<td>22 Douma B</td>
</tr>
<tr>
<td>3 Hausu M</td>
<td>23 Atem E</td>
</tr>
<tr>
<td>4 Atem L</td>
<td>24 Fossa P</td>
</tr>
<tr>
<td>5 Cier M</td>
<td>25 Houla H</td>
</tr>
<tr>
<td>6 Flanik B</td>
<td>26 Makan J</td>
</tr>
<tr>
<td>7 Mahong F</td>
<td>37 Houla B</td>
</tr>
<tr>
<td>8 Ngolle T</td>
<td>38 Houla E</td>
</tr>
<tr>
<td>9 German N</td>
<td>39 Nairiki G</td>
</tr>
<tr>
<td>10 Ague C</td>
<td>30 Haung E</td>
</tr>
<tr>
<td>11 Mavej M</td>
<td>31 Tifis M</td>
</tr>
<tr>
<td>12 Atem N</td>
<td>32 Peal N</td>
</tr>
<tr>
<td>13 Wawaru C</td>
<td>33 Kapue K</td>
</tr>
<tr>
<td>14 Zevegome M</td>
<td>34 Tifis E</td>
</tr>
<tr>
<td>15 Teblu M</td>
<td>35 Koudy F</td>
</tr>
<tr>
<td>16 Mahong F</td>
<td>36 Tifis E</td>
</tr>
<tr>
<td>17 Yaman H</td>
<td>37 Tiff K</td>
</tr>
<tr>
<td>18 Edeki J</td>
<td>38 Bowen E</td>
</tr>
<tr>
<td>19 Beke M</td>
<td>39 Tempa E</td>
</tr>
<tr>
<td>20 Haneck J</td>
<td>40 Bowen E</td>
</tr>
</tbody>
</table>

Selected households:

1 Isiere G
2 Mavej P
3 Meub P
4 Meub W
5 Beke M
6 Beke V
7 Beke V
8 Beke V
9 Beke V
10 Beke V
Systematic random sampling can be done with any list.

This is a part of the registry from a camp in south-eastern Nigeria. The original registry spreadsheet listed everyone living in the camp. Because we were doing a survey of adolescents, we used the field YOB (year of birth) to select only those camp residents who were 10-19 years of age at the time of the survey. The actual names in the column entitled NAME have been removed to preserve confidentiality.

Systematic random sampling can also do without a list. If the actual sampling units, such as houses or shelters, are arranged in order, you can count down the units in the field. Just calculate the sampling interval, choose a random number between 1 and the sampling interval, and then start counting the units from one end of the population.

The advantages of systematic sampling method over simple random sampling include:
It is easier to draw a sample and often easier to execute without mistakes. This is a particular advantage when the drawing is done in the field.

Intuitively, you might think that systematic sampling might be more precise than SRS. In effect it stratifies the population into n strata, consisting of the 1st k units, the 2nd k units, and so on. Thus, we might expect the systematic sample to be as precise as a stratified random sample with one unit per stratum. The difference is that with the systematic one the units occur at the same relative position in the stratum whereas with the stratified, the position in the stratum is determined separately by randomization within each stratum.

**Cluster Sampling**

In some instances the sampling unit consists of a group or cluster of smaller units that we call elements or subunits (these are the units of analysis for your study). There are two main reasons for the widespread application of cluster sampling. Although the first intention may be to use the elements as sampling units, it is found in many surveys that no reliable list of elements in the population is available and that it would be prohibitively expensive to construct such a list. In many countries there are no complete and updated lists of the people, the houses or the farms in any large geographical region.
Even when a list of individual houses is available, economic considerations may point to the choice of a larger cluster unit. For a given size of sample, a small unit usually gives more precise results than a large unit. For example a SRS of 600 houses covers a town more evenly than 20 city blocks containing an average of 30 houses apiece. But greater field costs are incurred in locating 600 houses and in traveling between them than in covering 20 city blocks. When cost is balanced against precision, the larger unit may prove superior.
Important things about cluster sampling:
  Most large scale surveys are done using cluster sampling;
  Clustering may be combined with stratification, typically by clustering within strata;
  In general, for a given sample size \( n \) cluster samples are less accurate than the other types of sampling in the sense that the parameters you estimate will have greater variability than an SRS, stratified random or systematic sample.

Non-probability Sampling
In a non-probability sample, some people have a greater, but unknown, chance than others of selection.

Social research is often conducted in situations where a researcher cannot select the kinds of probability samples used in large-scale social surveys. For example, say you wanted to study homelessness - there is no list of homeless individuals nor are you likely to create such a list. However, you need to get some kind of sample of respondents in order to conduct your research. To gather such a sample, you would likely use some form of non-probability sampling.

To reiterate, the primary difference between probability methods of sampling and non-probability methods is that in the latter you do not know the likelihood that any element of a population will be selected for study.

Advantages of Non-Probability Methods:
  - Cheaper
  - Used when sampling frame is not available
  - Useful when population is so widely dispersed that cluster sampling would not be efficient
  - Often used in exploratory studies, e.g. for hypothesis generation
  - Some research not interested in working out what proportion of population gives a particular response but rather in obtaining an idea of the range of responses on ideas that people have.
There are **five** primary types of non-probability sampling methods these are:

**Availability Sampling**

Availability sampling is a method of choosing subjects who are available or easy to find. This method is sometimes referred to as haphazard, accidental, or convenience sampling. The primary advantage of the method is that it is very easy to carry out, relative to other methods. A researcher can merely stand out on his/her favorite street corner or in his/her favorite tavern and hand out surveys. One place this used to show up often is in university courses. Years ago, researchers often would conduct surveys of students in their large lecture courses. For example, all students taking introductory sociology courses would have been given a survey and compelled to fill it out. There are some advantages to this design - it is easy to do, particularly with a captive audience, and in some schools you can attain a large number of interviews through this method.

The primary problem with availability sampling is that you can never be certain what population the participants in the study represent. The population is unknown, the method for selecting cases is haphazard, and the cases studied probably don’t represent any population you could come up with.

However, there are some situations in which this kind of design has advantages - for example, survey designers often want to have some people respond to their survey before it is given out in the "real" research setting as a way of making certain the questions make sense to respondents. For this purpose, availability sampling is not a bad way to get a group to take a survey, though in this case researchers care less about the specific responses given than whether the instrument is confusing or makes people feel bad.

Despite the known flaws with this design, it's remarkably common. Ask a provocative question, give telephone number and web site address ("Vote now at CNN.com), and announce results of poll. This method provides some form of statistical data on a current issue, but it is entirely unknown what population the result of such polls represents. At best, a researcher could make some conditional statement about people who are watching CNN at a particular point in time who cared enough about the issue in question to log on or call in.
Quota Sampling
Quota sampling is designed to overcome the most obvious flaw of availability sampling. Rather than taking just anyone, you set quotas to ensure that the sample you get represents certain characteristics in proportion to their prevalence in the population. Note that for this method, you have to know something about the characteristics of the population ahead of time. Say you want to make sure you have a sample proportional to the population in terms of gender - you have to know what percentage of the population is male and female, then collect sample until yours matches. Marketing studies are particularly fond of this form of research design.

The primary problem with this form of sampling is that even when we know that a quota sample is representative of the particular characteristics for which quotas have been set, we have no way of knowing if sample is representative in terms of any other characteristics. If we set quotas for gender and age, we are likely to attain a sample with good representativeness on age and gender, but one that may not be very representative in terms of income and education or other factors.

Moreover, because researchers can set quotas for only a small fraction of the characteristics relevant to a study quota sampling is really not much better than availability sampling. To reiterate, you must know the characteristics of the entire population to set quotas; otherwise there's not much point to setting up quotas. Finally, interviewers often introduce bias when allowed to self-select respondents, which is usually the case in this form of research. In choosing males 18-25, interviewers are more likely to choose those that are better-dressed, seem more approachable or less threatening. That may be understandable from a practical point of view, but it introduces bias into research findings.

Disadvantage of quota sampling: Interviewers choose who they like (within above criteria) and may therefore select those who are easiest to interview, so bias can result. Also, impossible to estimate accuracy (because not random sample)

Purposive Sampling
Purposive sampling is a sampling method in which elements are chosen based on purpose of the study. Purposive sampling may involve studying the entire population of some limited group (sociology faculty at
Columbia) or a subset of a population (Columbia faculty who have won Nobel Prizes). As with other non-probability sampling methods, purposive sampling does not produce a sample that is representative of a larger population, but it can be exactly what is needed in some cases - study of organization, community, or some other clearly defined and relatively limited group.

**Snowball Sampling**

Snowball sampling is a method in which a researcher identifies one member of some population of interest, speaks to him/her, then asks that person to identify others in the population that the researcher might speak to. This person is then asked to refer the researcher to yet another person, and so on.

Snowball sampling is very good for cases where members of a special population are difficult to locate. For example, several studies of Mexican migrants in Los Angeles have used snowball sampling to get respondents.

The method also has an interesting application to group membership - if you want to look at pattern of recruitment to a community organization over time, you might begin by interviewing fairly recent recruits, asking them who introduced them to the group. Then interview the people named, asking them who recruited them to the group.

The method creates a sample with questionable representativeness. A researcher is not sure who is in the sample. In effect snowball sampling often leads the researcher into a realm he/she knows little about. It can be difficult to determine how a sample compares to a larger population. Also, there's an issue of who respondents refer you to - friends refer to friends, less likely to refer to ones they don't like, fear, etc.

**Convenience Sampling**

A convenience sample is used when you simply stop anybody in the street who is prepared to stop, or when you wander round a business, a shop, a restaurant, a theatre or whatever, asking people you meet whether they will answer your questions. In other words, the sample comprises subjects who are simply available in a convenient way to the researcher. There is no randomness and the likelihood of bias is high. You can't draw any meaningful conclusions from the results you obtain.
However, this method is often the only feasible one, particularly for students or others with restricted time and resources, and can legitimately be used provided its limitations are clearly understood and stated.

Because it is an extremely haphazard approach, students are often tempted to use the word "random" when describing their sample where they have stopped people in the street, as they see it "at random". You should avoid using the word "random" when describing anything to do with sampling unless you are absolutely certain that you selected respondents from a sampling frame using truly random methods.

**Summary**

When undertaking any survey, it is essential that you obtain data from people that are representatives of the group that you are studying. Even with the perfect questionnaire (if such a thing exists), your survey data will only be regarded as useful if it is considered that your respondents are typical of the population as a whole. For this reason, an awareness of the principles of sampling is essential to the implementation of most methods of research, both quantitative and qualitative.

In this lecture, you have learnt the purpose of sampling, that is, to obtain study classes that are representative of the population under study. Also, we have seen the fundamental sampling techniques such as simple random, systematic, cluster, availability, quota and snowball sampling methods. All these save cost and highly skilled staff are employed to man the survey because of its size.

**Post – Test**

1. Discuss the advantages and disadvantages of stratified sampling compared with quota sampling.
2. Explain what is meant by a biased sample. Describe three forms of bias and state the precautions to be taken to minimize the risk of their occurrence.
3. Explain what is meant by a random sample and discuss its advantages and disadvantages.
4. Discuss the essentials of five methods of sampling.
5. What is quota sampling? Why would a social scientist frown at this sampling practice?
6. The management of a conglomerate is being advised to employ the stratified random scheme in a survey to study utilization of bank credits by the clients. Why would you support the scheme against a cheaper simple random alternative?
7. Write short notes on each of the following:
   i. Systematic sampling
   ii. Sampling frame
   iii. Random sampling
   iv. Stratification
   v. Probability sampling

References
LECTURE FOUR

Classification and Tabulation

Introduction
Classification and tabulation of data forms the basis for reducing and simplifying the details given in a mass of data into such a form that the main features may be brought out to make the assembled data easily understood.

The process of dividing the data into different groups (viz. classes) which are homogenous within but heterogeneous between them is called classification.

The main objective of the process of classification is to present the data, bringing out the points of similarity and classifying them into classes. The categories of classification may be by attributes or classification according to class interval if numerical/quantitative in nature. J. R. Hicks said “Classified and arranged facts speak of themselves, and narrated they are as dead as mutton”. It helps in understanding the salient features of the data and also the comparison with similar data.

Objectives
At the end of this lecture, readers are expected to discuss:
1. the meaning of classification;
2. types of classification;
3. meaning of tabulation and frequency distribution;
4. guideline for the construction of tables; and
5. how to group numeric data.
Pre-Test

1. Define Classification. Why the need for classification?
2. Enumerate the qualities of a good classification
3. Draw a dummy table showing the price per gallon and values of sales of a certain edible oil in 2005 and 2006 as recorded by a small scale industry.
4. Enumerate the guidelines for constructing table.
5. Define frequency and frequency distribution table
6. The following are monthly figures, x, of ratio of loans and advances to deposits by Nigeria Commercial Banks for the period January 2000 to April 2008. The figures are in percentages.

<table>
<thead>
<tr>
<th>Month</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>64.9</td>
<td>63.7</td>
<td>78.1</td>
<td>87.0</td>
<td>72.5</td>
<td>65.5</td>
<td>85.9</td>
<td>70.4</td>
<td></td>
</tr>
<tr>
<td>February</td>
<td>65.2</td>
<td>69.8</td>
<td>81.0</td>
<td>86.0</td>
<td>72.0</td>
<td>65.8</td>
<td>85.8</td>
<td>69.4</td>
<td></td>
</tr>
<tr>
<td>March</td>
<td>63.1</td>
<td>71.2</td>
<td>83.4</td>
<td>87.6</td>
<td>79.1</td>
<td>71.7</td>
<td>66.9</td>
<td>88.1</td>
<td>70.3</td>
</tr>
<tr>
<td>April</td>
<td>63.2</td>
<td>71.4</td>
<td>84.7</td>
<td>86.6</td>
<td>70.1</td>
<td>72.1</td>
<td>67.4</td>
<td>86.6</td>
<td>70.0</td>
</tr>
<tr>
<td>May</td>
<td>65.4</td>
<td>72.4</td>
<td>84.9</td>
<td>82.6</td>
<td>68.6</td>
<td>70.9</td>
<td>70.8</td>
<td>86.6</td>
<td></td>
</tr>
<tr>
<td>June</td>
<td>67.0</td>
<td>74.2</td>
<td>85.9</td>
<td>84.1</td>
<td>71.8</td>
<td>69.2</td>
<td>71.1</td>
<td>87.7</td>
<td></td>
</tr>
<tr>
<td>July</td>
<td>69.5</td>
<td>74.9</td>
<td>86.7</td>
<td>81.1</td>
<td>73.5</td>
<td>68.9</td>
<td>76.4</td>
<td>88.5</td>
<td></td>
</tr>
<tr>
<td>August</td>
<td>69.8</td>
<td>77.1</td>
<td>86.0</td>
<td>78.2</td>
<td>72.9</td>
<td>68.6</td>
<td>75.9</td>
<td>86.1</td>
<td></td>
</tr>
<tr>
<td>September</td>
<td>69.4</td>
<td>76.8</td>
<td>85.3</td>
<td>77.8</td>
<td>72.3</td>
<td>68.5</td>
<td>78.4</td>
<td>83.2</td>
<td></td>
</tr>
<tr>
<td>October</td>
<td>67.8</td>
<td>78.8</td>
<td>86.6</td>
<td>79.3</td>
<td>72.4</td>
<td>68.0</td>
<td>84.1</td>
<td>81.6</td>
<td></td>
</tr>
<tr>
<td>November</td>
<td>66.2</td>
<td>83.4</td>
<td>87.1</td>
<td>78.2</td>
<td>73.0</td>
<td>68.1</td>
<td>87.6</td>
<td>80.9</td>
<td></td>
</tr>
<tr>
<td>December</td>
<td>63.7</td>
<td>80.6</td>
<td>85.5</td>
<td>79.6</td>
<td>73.1</td>
<td>69.3</td>
<td>86.7</td>
<td>75.9</td>
<td></td>
</tr>
</tbody>
</table>

a. Determine the range and the midrange of the data.
b. Summarize the data into a frequency distribution using class groupings 60 – 64, 65 – 69, 70 – 74, etc.

CONTENT

What is Classification?

Classification is a process of arranging observations into logically, meaningful, useful categories in accordance with the nature of property understudy.
The idea is to group like things together to facilitate comparison of likes with like and also to reduce a large number of data to a form suitable for statistical analysis and processing. Ideally, a group of class must be homogenous, that is, it should include all items and only those items with definite characteristic of data.

a. **Qualitative classification**: when people or things are sorted in groups, each possessing some attributes that cannot be expressed numerically e.g. employers in an establishment can be grouped as union or non-union, high, middle and lower social classes, gender-male or female, divorced married or single.

b. **Quantitative classification**: when item varies in respect of some measurable characteristic. Most quantitative classification form frequency distribution as we shall see very soon.

c. **Geographical classification**: is a classification by geographical location and it is very necessary for administrative purposes. It is also a form of qualitative classification.

d. **Chronological classification**: chronological data or time series shows figure concerning a particular phenomenon at various specified time e.g. birth and death rate listed for each year, inflation rate listed for period of time, etc.

**Qualities of a good classification**

The following are the qualities of a good classification:

1. The items grouped must be homogenous.
2. The classes must be exhaustive, that is, each item must have a class to which it belongs
3. The classes must be exclusive, that is, each item must belong to only one class.

**What is Tabulation?**

Tabular presentation is an orderly arrangement of numerical information in columns and rows in order to present statistical information in a concise and orderly fashion. After the collection of data, the analyst first task is to reduce and simplify the detail into such a form that the salient features of the survey may be brought out. This will facilitate the interpretation of the assembled data. This procedure is known as classifying and tabulating the
data, which is, extracting from the individual questionnaire or schedule the answer to each question and entering the replies on separate summary sheets. This form the basis of reducing and simplifying the details given in a mass of data into such a form that the main features may be brought out to make the assembled data easily understood. This is done in a statistical table where the data may be arranged in columns and or rows.

Statistical tables therefore are the instrument used to organize and present the data obtained from a survey. These tables should be designed to meet exact technical specifications for tabular presentation for tabular presentation. It is very important to maintain simplicity in the design of a statistical table. There are four types of tables, viz, text table or summary, working table, general or reference table and information or classifying tables.

**Tables**

Tables are designed to organize and present numerical information so that it can be understood and interpreted by all. The purpose of a table is to be able to study the relationship between variables with their totals.

**Guidelines for the Construction of Tables**

1. Table should be simple and unambiguous.
2. It must be easily interpreted.
3. It should present the data clearly, highlighting important details.
4. It should save space but attractively designed.
5. The table number and title should be given.
6. Units of the measurement should be clearly stated along the titles or headings.
7. Abbreviations and symbols should be avoided as much as possible.
8. Sources of the data should be given at the bottom of the data.
9. Where the data are classified in frequency distribution, there should be no overlapping of limits of the successive classes.

**Frequency Distribution**

Frequency distribution is a tabular arrangement of data by classes together with the corresponding class frequencies.
The idea is that, the result of survey is the bulk of raw-data obtained. It is said to be raw when it is still in the crude form and is yet to be processed. Thus it has neither been edited nor classified. What exists in this circumstance, therefore, is a heap of disorganized facts which have little or no meaning and which cannot be used directly for any reasonable analysis. This is useless and meaningless unless these are condensed, that is, distributing into classes or categorize and to determine the number of items or individuals belonging to each class called the class frequency.

The number of times a value appears in the listing is referred to as its frequency.

Raw data are collected data which have not been organized numerically. An array is an arrangement of raw numerical data in ascending or descending order of magnitude.

The following are steps in forming a frequency distribution table from raw data.

1. Find the range of the data, that is, the difference between highest and smallest observations.

2. Determine the number of classes by dividing the range into a convenient number of class intervals having the same size. The number of class intervals is usually taken between 5 and 20, that is, $5 \leq x \leq 20$ depending on the data.

3. The length of a class or class interval is determined by the number of classes and the range of the data, that is, we divide the range by the number of classes chosen in (2). But ideally we should have the number that is easy to work with e.g. 1, 2, 5, 20, etc. Class intervals are also chosen so that the class marks or midpoints coincide with actual observed data. This tends to lessen the so called grouping error involved in further statistical analysis.

4. Determine the number of observations falling into each class interval, that is, find the class frequencies. This is best done by using tally or score sheet.

5. Classes should not overlap and no gap exists.
Simple illustration of the above is given below:

**Example 1:** The ages in years of fifty members of House of Representatives are as follows:

<table>
<thead>
<tr>
<th>Ages in years</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>IIII</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>IIIIII</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>IIIII</td>
<td>7</td>
</tr>
<tr>
<td>31</td>
<td>IIII</td>
<td>6</td>
</tr>
<tr>
<td>32</td>
<td>IIIII</td>
<td>7</td>
</tr>
<tr>
<td>33</td>
<td>III</td>
<td>4</td>
</tr>
<tr>
<td>34</td>
<td>IIII</td>
<td>6</td>
</tr>
<tr>
<td>35</td>
<td>III</td>
<td>4</td>
</tr>
</tbody>
</table>

Prepare a frequency table for the above data.

**Solution**

We can see that the data is between 28 and 35, hence we may not be able to group this into interval classes because we may not be able to realize the satisfactory number of classes to make it a good frequency table.
Table 1

Example 2: The data below are the monthly wages of 100 workers in (₦ ‘000) in a certain firm.

<table>
<thead>
<tr>
<th>88</th>
<th>82</th>
<th>96</th>
<th>102</th>
<th>104</th>
<th>106</th>
<th>104</th>
<th>24</th>
<th>26</th>
<th>29</th>
<th>86</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>60</td>
<td>23</td>
<td>24</td>
<td>39</td>
<td>48</td>
<td>46</td>
<td>33</td>
<td>36</td>
<td>39</td>
<td>78</td>
<td>67</td>
<td>82</td>
</tr>
<tr>
<td>32</td>
<td>67</td>
<td>27</td>
<td>24</td>
<td>26</td>
<td>27</td>
<td>30</td>
<td>36</td>
<td>37</td>
<td>49</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>83</td>
<td>99</td>
<td>68</td>
<td>28</td>
<td>55</td>
<td>54</td>
<td>26</td>
<td>29</td>
<td>30</td>
<td>40</td>
<td>46</td>
<td>44</td>
</tr>
<tr>
<td>99</td>
<td>84</td>
<td>36</td>
<td>51</td>
<td>86</td>
<td>88</td>
<td>87</td>
<td>29</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>66</td>
</tr>
<tr>
<td>45</td>
<td>23</td>
<td>26</td>
<td>46</td>
<td>46</td>
<td>96</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>101</td>
<td>103</td>
<td>106</td>
</tr>
<tr>
<td>107</td>
<td>46</td>
<td>48</td>
<td>49</td>
<td>48</td>
<td>94</td>
<td>55</td>
<td>56</td>
<td>59</td>
<td>60</td>
<td>70</td>
<td>72</td>
</tr>
<tr>
<td>76</td>
<td>79</td>
<td>80</td>
<td>50</td>
<td>49</td>
<td>93</td>
<td>86</td>
<td>54</td>
<td>83</td>
<td>89</td>
<td>90</td>
<td>94</td>
</tr>
<tr>
<td>96</td>
<td>99</td>
<td>102</td>
<td>46</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Prepare a suitable frequency table for the data.

Solution

The biggest/Highest (H) value is 106 and the smallest value (S) is 23, therefore the range is \( H - L = 106 - 23 = 83 \). There are 100 items, hence by Sturges’ rule number of classes should be

\[
1 + 33\log_{10}100 = 7.6 \approx 8.
\]

For 8 groups the new class interval will be \( 83/8 = 10.4 \). Therefore, a class interval of 10 will be suitable. The lower class limit of the first class will therefore be 20 (that is the lower class limit should also be a multiple of 5 or 10. Hence, the first group is taken as 20 and below 30. Thus the tally bar chart and the required distribution will be as below:
<table>
<thead>
<tr>
<th>Daily Wages (N’000)</th>
<th>Tally</th>
<th>No. of Workers (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 and below 30</td>
<td>### ### /</td>
<td>13</td>
</tr>
<tr>
<td>30 and below 40</td>
<td>### /</td>
<td>11</td>
</tr>
<tr>
<td>40 and below 50</td>
<td>### ### ###</td>
<td>18</td>
</tr>
<tr>
<td>50 and below 60</td>
<td>###</td>
<td>10</td>
</tr>
<tr>
<td>60 and below 70</td>
<td>### /</td>
<td>6</td>
</tr>
<tr>
<td>70 and below 80</td>
<td>###</td>
<td>5</td>
</tr>
<tr>
<td>80 and below 90</td>
<td>### ### /</td>
<td>14</td>
</tr>
<tr>
<td>90 and below 100</td>
<td>### /</td>
<td>12</td>
</tr>
<tr>
<td>100 and below 110</td>
<td>### /</td>
<td>11</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>100</td>
</tr>
</tbody>
</table>

**Table 2**
We want to see the third example to enable us understand some terms as we go along in the study of statistics.

**Example 3:** The marks scored by 50 students in a particular course are as follows:

<table>
<thead>
<tr>
<th>Marks</th>
<th>38</th>
<th>74</th>
<th>28</th>
<th>32</th>
<th>10</th>
<th>31</th>
<th>49</th>
<th>34</th>
<th>50</th>
<th>19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marks</td>
<td>30</td>
<td>92</td>
<td>50</td>
<td>42</td>
<td>38</td>
<td>64</td>
<td>24</td>
<td>65</td>
<td>9</td>
<td>77</td>
</tr>
<tr>
<td>Marks</td>
<td>18</td>
<td>35</td>
<td>12</td>
<td>87</td>
<td>41</td>
<td>27</td>
<td>8</td>
<td>90</td>
<td>22</td>
<td>21</td>
</tr>
<tr>
<td>Marks</td>
<td>42</td>
<td>43</td>
<td>52</td>
<td>59</td>
<td>72</td>
<td>70</td>
<td>90</td>
<td>91</td>
<td>29</td>
<td>28</td>
</tr>
<tr>
<td>Marks</td>
<td>45</td>
<td>46</td>
<td>43</td>
<td>55</td>
<td>56</td>
<td>59</td>
<td>44</td>
<td>53</td>
<td>47</td>
<td>75</td>
</tr>
</tbody>
</table>
Prepare a frequency table using class intervals 1 – 20, 21 – 40, 41 – 60, etc.

Solution

The class intervals and classes have been given, we can now construct the table

<table>
<thead>
<tr>
<th>Scores</th>
<th>Tally</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 20</td>
<td>### /</td>
<td>6</td>
</tr>
<tr>
<td>21 – 40</td>
<td>### ### ///</td>
<td>14</td>
</tr>
<tr>
<td>41 – 60</td>
<td>### ### ### ///</td>
<td>19</td>
</tr>
<tr>
<td>61 – 80</td>
<td>### /</td>
<td>6</td>
</tr>
<tr>
<td>81 - 100</td>
<td>###</td>
<td>5</td>
</tr>
</tbody>
</table>

Table 3

Explanation of some terminologies under classification and tabulation

1. **Class intervals**: is the description of the class that has both the upper class limit and lower class limit. A class that does not have either an upper and lower limit is called an open - ended class.

2. **Class limits**: A class is formed within the two values. These values are known as the class limits of that class. The lower value is called the lower class limits and is denoted \( l_1 \) while the higher value is called the upper limit of the class and is denoted \( l_2 \). In the table 3 above, the first class is 1 – 20, the lower class limit is 1 while the upper class limit is 20.
3. **Class mark**: is defined as the midpoint of a class interval. It is sometimes described as the arithmetic average of the two class limits (that is the lower limit and the upper limit). It is computed by adding the lower and upper class limits of a class interval and then dividing by 2.

\[
\text{Class mark} = \frac{\text{lower class limit} + \text{upper class limit}}{2}
\]

The midpoint of the first class in table 3 above is \( \frac{1 + 20}{2} = 10.5 \)

4. **Class size**: is the difference between the upper and lower class boundaries is called the class size or the magnitude or length or width of a class and is denoted “c”.

5. **Class boundaries**: Weights are recorded to the nearest Kg. The class interval being used in table 3 above 1 – 20 includes all measurements from 0.5 to 20.5, the variable being a continuous one are called the class boundaries. The smaller being 0.5 and larger one is 20.5. The class boundary has a lot of advantages over class limits. You can determine the true class size from the difference between the upper and lower class boundaries. Some charts or graph cannot be plotted successfully unless the data are transformed to class boundaries, e.g Histogram and ogive cannot be plotted unless the data are adjusted to boundaries. For example, in table 3, the first class, the boundaries are obtained by deducting 0.5 from the lower class limit and 0.5 added to the upper class boundaries. It can also be described as a point that represents the halfway, or dividing, point between successive classes. It is a point where the lower class ends and the higher class begins

6. **Class frequency**: is the number of values that fall in a class represents the frequency of the class and is called the class frequency.

7. **Relative frequency**: the relative frequency of a class is the frequency of the class divided by the total frequency (the sum of all the frequencies of a given table).
8. **Cumulative frequency**: is the total frequency of all values less than or equal to the upper class boundary of a given class interval.

9. **Relative cumulative frequency**: The relative cumulative frequency, or percentage cumulative frequency is the cumulative frequency divided by the total frequency.

**Example**: The following data give the amounts (in N'000) on feeding by forty households in a month.

<table>
<thead>
<tr>
<th>32</th>
<th>22</th>
<th>19</th>
<th>18</th>
<th>43</th>
<th>42</th>
<th>40</th>
<th>43</th>
<th>18</th>
<th>21</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>26</td>
<td>22</td>
<td>25</td>
<td>47</td>
<td>40</td>
<td>26</td>
<td>32</td>
<td>34</td>
<td>22</td>
</tr>
<tr>
<td>35</td>
<td>38</td>
<td>34</td>
<td>41</td>
<td>36</td>
<td>25</td>
<td>22</td>
<td>45</td>
<td>48</td>
<td>26</td>
</tr>
<tr>
<td>28</td>
<td>38</td>
<td>35</td>
<td>19</td>
<td>47</td>
<td>28</td>
<td>26</td>
<td>35</td>
<td>38</td>
<td>35</td>
</tr>
</tbody>
</table>

Construct a frequency distribution using seven classes. Hence compute

i. The class mark for each class
ii. The class boundaries for each class
iii. The cumulative frequency of the distribution
iv. The relative frequencies of the classes
v. The relative cumulative frequency

**Solution**

The smallest value is 18 and the largest value is 48, therefore, the range is

\[ 48 - 18 = 30 \]

Since we are asked to form seven classes, the approximate length of a class interval is \( 30/7 \approx 4.29 \approx 5 \). We then take the length of class interval as 5, a convenient starting point is 15, thus the first class would be 15 – 19 follow by 20 – 24, 25 – 27, …… 45 – 49. The resulting frequency distribution is as given below:
Expenditure on feeding (N’000) | Tally | No of households
---|---|---
15 – 19 | //// | 4
20 – 24 | /// | 5
25 – 29 | /// /// | 8
30 – 34 | /// | 5
35 – 39 | /// /// | 8
40 – 44 | /// / | 6
45 - 49 | /// | 4
Total | | 40

**Figure:** frequency distribution of expenditure on feeding for some selected households.

<table>
<thead>
<tr>
<th>Expenditures</th>
<th>No of Household</th>
<th>Class Mark</th>
<th>Class Boundaries</th>
<th>Cumulative Frequency</th>
<th>Relative Frequency</th>
<th>Relative Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 – 19</td>
<td>4</td>
<td>17</td>
<td>14.5 – 19.5</td>
<td>4</td>
<td>1/10</td>
<td>10%</td>
</tr>
<tr>
<td>20 – 24</td>
<td>5</td>
<td>22</td>
<td>19.5 – 24.5</td>
<td>9</td>
<td>1/8</td>
<td>22.5%</td>
</tr>
<tr>
<td>25 – 29</td>
<td>8</td>
<td>27</td>
<td>24.5 – 29.5</td>
<td>17</td>
<td>1/5</td>
<td>42.5%</td>
</tr>
<tr>
<td>30 – 34</td>
<td>5</td>
<td>32</td>
<td>29.5 – 34.5</td>
<td>22</td>
<td>1/8</td>
<td>55%</td>
</tr>
<tr>
<td>35 – 39</td>
<td>8</td>
<td>37</td>
<td>34.5 – 39.5</td>
<td>30</td>
<td>1/5</td>
<td>75%</td>
</tr>
<tr>
<td>40 – 44</td>
<td>6</td>
<td>42</td>
<td>39.5 – 44.5</td>
<td>36</td>
<td>3/20</td>
<td>90%</td>
</tr>
<tr>
<td>45 – 49</td>
<td>4</td>
<td>47</td>
<td>44.5 – 49.5</td>
<td>40</td>
<td>1/10</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Summary**

Data obtained from statistical enquiry is often so numerous that one’s mind can hardly comprehend its significance in the form that it is shown. Therefore it becomes, very necessary to tabulate and summarize the data to an easily manageable form. In doing so we may overlook its details. But this is not a serious loss because we are not interested in an individual but in the properties of aggregates. To a layman, presentation of the raw data in the form of tables or diagrams is always more effective. Tabulation is therefore a process of condensation of the data.
Post-Test

1. What is classification? List out different types of classification.
2. What is tabulation? What are the purposes of tabulation?
3. State the requirements for constructing a good table.
4. Define the following:
   i. Class limits
   ii. Class intervals
   iii. Class boundary
   iv. Class frequency.
5. Define frequency distribution. State three important factors which should be considered in constructing a frequency table.
6. The following are the monthly wages of workers in a given establishment.

<table>
<thead>
<tr>
<th>88</th>
<th>82</th>
<th>96</th>
<th>102</th>
<th>104</th>
<th>106</th>
<th>104</th>
<th>24</th>
<th>26</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>86</td>
<td>36</td>
<td>60</td>
<td>23</td>
<td>24</td>
<td>39</td>
<td>48</td>
<td>46</td>
</tr>
<tr>
<td>33</td>
<td>36</td>
<td>39</td>
<td>78</td>
<td>67</td>
<td>82</td>
<td>32</td>
<td>67</td>
<td>27</td>
</tr>
<tr>
<td>24</td>
<td>26</td>
<td>27</td>
<td>30</td>
<td>36</td>
<td>37</td>
<td>49</td>
<td>50</td>
<td>56</td>
</tr>
<tr>
<td>83</td>
<td>99</td>
<td>68</td>
<td>28</td>
<td>55</td>
<td>54</td>
<td>26</td>
<td>29</td>
<td>30</td>
</tr>
<tr>
<td>40</td>
<td>46</td>
<td>44</td>
<td>99</td>
<td>84</td>
<td>36</td>
<td>51</td>
<td>86</td>
<td>88</td>
</tr>
<tr>
<td>87</td>
<td>29</td>
<td>40</td>
<td>40</td>
<td>40</td>
<td>66</td>
<td>45</td>
<td>23</td>
<td>26</td>
</tr>
<tr>
<td>46</td>
<td>46</td>
<td>96</td>
<td>99</td>
<td>100</td>
<td>100</td>
<td>101</td>
<td>103</td>
<td>106</td>
</tr>
<tr>
<td>107</td>
<td>46</td>
<td>48</td>
<td>49</td>
<td>48</td>
<td>94</td>
<td>55</td>
<td>56</td>
<td>59</td>
</tr>
<tr>
<td>60</td>
<td>70</td>
<td>72</td>
<td>76</td>
<td>79</td>
<td>80</td>
<td>50</td>
<td>49</td>
<td>93</td>
</tr>
<tr>
<td>86</td>
<td>54</td>
<td>83</td>
<td>89</td>
<td>90</td>
<td>94</td>
<td>96</td>
<td>99</td>
<td>102</td>
</tr>
</tbody>
</table>

Determine the range, class interval for the distribution above (make the number of classes to be 9). Use tally method to classify the above data into classes.

7. Compute the class marks and class boundaries for the above distribution in (6).

8. The distribution of time spent (in minutes) in a bank by 100 customers are as follows:
<table>
<thead>
<tr>
<th>Data</th>
<th>22.7</th>
<th>25.9</th>
<th>34.2</th>
<th>13.4</th>
<th>26.4</th>
<th>29.5</th>
<th>30.5</th>
<th>18.2</th>
<th>24.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>28.4</td>
<td>33.6</td>
<td>41.0</td>
<td>16.8</td>
<td>28.5</td>
<td>28.3</td>
<td>18.7</td>
<td>15.9</td>
<td>23.2</td>
</tr>
<tr>
<td>Data</td>
<td>19.4</td>
<td>33.8</td>
<td>40.5</td>
<td>21.0</td>
<td>31.6</td>
<td>22.5</td>
<td>31.5</td>
<td>26.5</td>
<td>31.8</td>
</tr>
<tr>
<td>Data</td>
<td>21.5</td>
<td>26.3</td>
<td>28.4</td>
<td>35.5</td>
<td>25.5</td>
<td>39.5</td>
<td>23.1</td>
<td>26.5</td>
<td>24.0</td>
</tr>
<tr>
<td>Data</td>
<td>28.3</td>
<td>11.0</td>
<td>29.1</td>
<td>36.4</td>
<td>26.4</td>
<td>43.6</td>
<td>20.6</td>
<td>28.3</td>
<td>29.4</td>
</tr>
<tr>
<td>Data</td>
<td>25.7</td>
<td>26.9</td>
<td>30.8</td>
<td>27.5</td>
<td>32.5</td>
<td>33.4</td>
<td>24.2</td>
<td>33.1</td>
<td>29.4</td>
</tr>
<tr>
<td>Data</td>
<td>25.8</td>
<td>17.5</td>
<td>26.6</td>
<td>35.6</td>
<td>29.0</td>
<td>27.4</td>
<td>10.8</td>
<td>26.2</td>
<td>36.7</td>
</tr>
<tr>
<td>Data</td>
<td>44.5</td>
<td>31.4</td>
<td>24.4</td>
<td>25.8</td>
<td>32.8</td>
<td>22.1</td>
<td>34.4</td>
<td>25.9</td>
<td>14.5</td>
</tr>
<tr>
<td>Data</td>
<td>27.4</td>
<td>28.7</td>
<td>43.8</td>
<td>15.5</td>
<td>24.3</td>
<td>37.3</td>
<td>27.3</td>
<td>16.5</td>
<td>42.3</td>
</tr>
<tr>
<td>Data</td>
<td>38.4</td>
<td>34.0</td>
<td>27.2</td>
<td>25.6</td>
<td>34.5</td>
<td>23.5</td>
<td>34.4</td>
<td>27.8</td>
<td>33.6</td>
</tr>
<tr>
<td>Data</td>
<td>29.4</td>
<td>26.4</td>
<td>24.5</td>
<td>37.5</td>
<td>29.5</td>
<td>10.6</td>
<td>27.6</td>
<td>19.5</td>
<td>38.0</td>
</tr>
<tr>
<td>Data</td>
<td>19.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Construct a frequency distribution using eight classes. Hence compute:

i. The class mark for each class
ii. The class boundaries for each class
iii. The cumulative frequency of the distribution
iv. The relative frequencies of the classes
v. The relative cumulative frequency

References


LECTURE FIVE

Presentation of Data (Charts and Graphs)

Introduction
In the last lecture we have seen how to condense the mass of data by the method of classification and tabulation. It is not always easy for people to understand figures. Apart from the fact that too many figures are often confusing. One of the most convincing and appealing ways in which statistical results may be represented is through graphs and diagrams. It is for this reason that diagrams are often-used by government agencies, businessmen, newspapers, journals, and biomedical personnel for advertising and educating people.

Objectives
Students are expected at the end of this lecture to master the following:
1. compare tabular and diagrammatic presentation;
2. the advantages and disadvantages of graphical presentation; and
3. different types of charts and diagram and the type of data they are representing.

Pre-Test
1. List any seven types of statistical diagrams.
2. The sales records at a supermarket over a period of two years were as follows:

<table>
<thead>
<tr>
<th>Months</th>
<th>Year I (N’000)</th>
<th>Year II (N’000)</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>February</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>March</td>
<td>8</td>
<td>9</td>
</tr>
</tbody>
</table>
Illustrate the progress chart of the sales by means of Z-curve.

3. a. Mention five likely application areas of a Lorenz curve.
   b. The table below shows the figures from an annual report of manufacturing industries of seven industrial cities of a country in 1998.

<table>
<thead>
<tr>
<th>Net Output (N’million)</th>
<th>Number of Establishments</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td>48</td>
</tr>
<tr>
<td>23</td>
<td>42</td>
</tr>
<tr>
<td>28</td>
<td>21</td>
</tr>
<tr>
<td>32</td>
<td>26</td>
</tr>
<tr>
<td>37</td>
<td>36</td>
</tr>
<tr>
<td>50</td>
<td>16</td>
</tr>
<tr>
<td>74</td>
<td>23</td>
</tr>
</tbody>
</table>

Analyze this table by means of Lorenz curve and explain what this curve shows.

4. i. State the main purpose of the Gantt Progress Chart.
   ii. The budgeted revenue and the actual income from the records of the University of Ibadan within the last six years are given as follows:
### Yearly Budget Analysis

<table>
<thead>
<tr>
<th>Year</th>
<th>Budget (N billion)</th>
<th>Actual (N billion)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>3.2</td>
<td>2.3</td>
</tr>
<tr>
<td>2003</td>
<td>3.5</td>
<td>3.3</td>
</tr>
<tr>
<td>2004</td>
<td>3.7</td>
<td>3.8</td>
</tr>
<tr>
<td>2005</td>
<td>3.9</td>
<td>3.1</td>
</tr>
<tr>
<td>2006</td>
<td>4.3</td>
<td>3.5</td>
</tr>
<tr>
<td>2007</td>
<td>5.2</td>
<td>4.0</td>
</tr>
</tbody>
</table>

i. Draw a Gantt progress chart for the University.


### CONTENT

#### Charts and Graph

Statistical data can often be presented in chart form, diagram or graph. This enables relationships and trends and comparisons to be grasped more readily. They are used to organize, condense, communicate, expose ideas and concepts of the usual delineation.

Comparison between Tabular and Diagrammatic Presentation

| 1. Diagrams and graphs are meant for a layman | 1. Tables are meant for statistician for the purpose of further analysis |
| 2. Diagrams give only an approximate idea. | 2. Tables contain precise figures. Exact values can be read from tables. |
| 3. Diagrams can be more easily compared, and can be interpreted by a layman. | 3. Comparison and interpretations of tables can only be done by statisticians and it is a difficult task. |
| 4. Diagrams and graphs cannot present much information | 4. Tables can present more information |
| 5. Diagrams are more attractive and have a visual appeal. | 5. Tables are dry for a layman (may be attractive to a statistician) |
Advantages of Graphical Presentation

1. **Quicker communication**: Particularly where clarity is essential
2. **More Dynamic**: Graphic carries greater weight through its ability to demonstrate.
3. **More Revealing**: Brings out relationship that might otherwise be undetected
4. **More stimulating**: It provides a contrast on the written or spoken word.
5. **More memorable**: Projects a distinctive image of the situation
6. **Audience participation**: It offers a good speaker the opportunity to press home his argument.
7. **Accuracy cross-check**: It spotlights immediately some odd placings as a result of some error in basic data.

Disadvantages

1. **Message limitation**: Single graph can only communicate a limited number of messages if it is to avoid risk of confusion.
2. **Cost**: It is costly
3. **Time**: It is time consuming
4. **Personal preference**: Certain people are suspicious of graph.

General Principle of Graph construction

1. The correct impression must be given
2. The graph must have a clear, concise and suitable title without damaging clarity.
3. The independent variable should always be placed on the horizontal axis.
4. The vertical scale should always start at zero.
5. Axes should be clearly labeled
6. Curves must be distinct
7. The graph must not be overcrowded with curves
8. The source of the data must always be given.
9. The diagrams should be simple.
10. An index, explaining different lines, shades and colours should be given.
11. Diagrams should be absolutely neat and clean.

Williams play fair (1759 – 1823) – said in his commercial and political Atlas of 1801 that “Information take days through data presented in tables, but can be obtained in minutes through diagrams.”

Cold figures are uninspiring to most people. Diagrams help us to see pattern and shape of any complex idea, just as a map gives a bird’s eye view of wide sketch of a country, so diagrams help us to visualize the whole meaning of a numerical complex at single glance. Diagrams register a meaningful impression almost before you think……” (Moroney

Types of Charts, Diagrams and Graph

These are of four categories:

a. Graphical presentation of frequency distribution
   i. Histogram
   ii. Frequency polygon
   iii. Cumulative frequency polygon

b. Diagrams use in displaying non-numeric data.
   i. Pictograms
   ii. Simple bar charts
   iii. Pie charts

c. Diagrams /graphs for miscellaneous data
   i. Component, percentage and multiple bar charts
   ii. Multiple pie charts
   iii. Z – charts
   iv. Gantt charts, semi-logarithmic graphs
   v. Ratio scale
   vi. Lorenz curves
   vii. Gini – curves
   viii. Pareto curves

d. Diagrams for displaying time series data
   i. Line diagrams
   ii. Simple bar charts
a. **Graphical presentation of frequency distribution**

The most common and convenient ways to present a frequency table graphically are by means of histogram, frequency polygon and cumulative frequency polygon (ogive) and frequency curve. These are called graphs. A graph is a visual representation of data by a continuous curve on a squared (graph) paper. Like diagrams, graphs are also attractive, and eye-catching, giving a bird’s eye-view of data and revealing their inner pattern.

i. **Histogram:** It is defined as a pictorial representation of a grouped frequency distribution by means of adjacent rectangles, whose areas are proportional to the frequencies. It is a graphical representation of a frequency distribution in which class frequencies are plotted against class boundaries. The reason for the class boundaries is that the histogram bars share common boundaries, that is, the classes overlap. The class frequency is then represented by vertical bars forming a rectangular block whose area is proportional to the frequency and width is equal to the length of the class.

In case of classes having unequal widths, rectangles too stand on unequal widths (bases). for open-classes, histogram is constructed after making certain assumptions. As the rectangles are adjacent leaving no gaps, the class-intervals become of the inclusive type, adjustment is necessary for end points only.

**Example:** Consider the table below:

<table>
<thead>
<tr>
<th>Scores</th>
<th>Frequency</th>
<th>Class boundaries</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 20</td>
<td>6</td>
<td>0.5 – 20.5</td>
</tr>
<tr>
<td>21 – 40</td>
<td>8</td>
<td>20.5 – 40.5</td>
</tr>
<tr>
<td>41 – 60</td>
<td>14</td>
<td>40.5 – 60.5</td>
</tr>
<tr>
<td>61 – 80</td>
<td>6</td>
<td>60.5 – 80.5</td>
</tr>
<tr>
<td>81 – 100</td>
<td>4</td>
<td>80.5 – 100.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
<td></td>
</tr>
</tbody>
</table>
ii. **Frequency Polygon**: A frequency polygon is a graphical representation of a frequency distribution in which class frequencies are plotted against the class marks. These are then joined by straight line segments. The class mark is

\[
\text{Class Mark} = \frac{\text{Lower class limit} + \text{Upper class limit}}{2}
\]

The frequency polygon can also be drawn by joining the mid-points of the bars in histogram. The polygon is closed at the base by extending it on both its sides (ends) to the midpoints of two hypothetical classes, at the extremes of the distribution, with zero frequencies.
Example: From the above data above, we now compute the class marks.

<table>
<thead>
<tr>
<th>Class mark</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10.5</td>
<td>6</td>
</tr>
<tr>
<td>30.5</td>
<td>8</td>
</tr>
<tr>
<td>50.5</td>
<td>14</td>
</tr>
<tr>
<td>70.5</td>
<td>6</td>
</tr>
<tr>
<td>90.5</td>
<td>4</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

Figure: Frequency Polygon
iii. **Cumulative Frequency Curve (Ogive):** A cumulative frequency polygon is also a graphical representation in which upper class boundaries are plotted against cumulative frequencies. Cumulative frequencies can be computed thus.

<table>
<thead>
<tr>
<th>Class boundaries</th>
<th>Frequency</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 0.5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Less than 20.5</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Less than 40.5</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>Less than 60.5</td>
<td>16</td>
<td>30</td>
</tr>
<tr>
<td>Less than 80.5</td>
<td>6</td>
<td>36</td>
</tr>
<tr>
<td>Less than 100.5</td>
<td>4</td>
<td>40</td>
</tr>
</tbody>
</table>
Diagrams Use in Displaying Non-numeric Data.

a. Pictograms
b. Simple bar charts
c. Pie charts

a. Pictograms

Pictogram is a chart which represents the magnitude of numerical value by using only simple descriptive pictures. They are used to make visual comparison in exactly the same way as a bar chart, and, in effect the numbers of symbols correspond to the length of a bar chart.
**Example:** The following represent the number of students in the under listed faculties:

- **Arts** – 2500
- **Science** – 3000
- **Technology** – 1500
- **Social sciences** – 3500
- **Law** – 1000

Draw pictograms to represent all these figures.

**Solution**

**Arts**

![Pictogram for Arts](image)

1000 1000 500

**Science**

![Pictogram for Science](image)

**Technology**

![Pictogram for Technology](image)
Social Sciences

![Smiley faces representing different ratings.]

Law

![A single smiley face.]

Key

![Smiley faces with different sizes.]

1000 students  500 students

**Advantage**
It is easy to understand for a non-sophisticated audience.

**Disadvantage**
It can be awkward to construct if complex symbols are used.

**Simple Bar Chart**
A simple bar chart is a chart consisting of a set of non-joining bars. A separate bar for each class is drawn to a height proportional to the class frequency or the values that they represent. The widths of the bars drawn for each class are always the same, and desired, each bar can be shaded or coloured differently. The bars can be horizontally or vertically oriented; vertical bars are more popular. Sometimes a stretched graphic is used instead of a solid bar. The bar chart enables magnitude to be compared.
Advantages
1. A dramatic and appealing way of presenting data.
2. It is good for comparing classes in relative terms.
3. Simple bar diagrams are very popular in practice.
4. It makes relationship easy to study for readers.

Disadvantages
1. Compilation is laborious
2. It can be untidy if there are many classes.

Example
The following table lists the number of seats won by each party in the Upper House of Assembly:

PDP – 56  
ANPP – 27  
AC -  12  
Labour – 8  
APGA – 3  
AD - 2

Represent the information on simple bar chart

Solution
Pie Chart

A pie chart (or a circle graph) is a popular device for showing how an aggregate is divided into its principal components. It is a circular chart divided into sectors, illustrating relative magnitudes or frequencies or percents. Like the component bar chart, it shows the relationship of parts to the whole. It is a convenient way of showing the sizes of component figures in proportion to each other and to the overall total. In component bar chart lengths of bars are compared whereas in a pie chart, areas of segments are compared. Because it is difficult to compare areas visually pie charts are an inferior form of presentation. In a pie chart, the arc length of each sector (and consequently its central angle and area), is proportional to the quantity it represents. The angles are measured in order to obtain the frequencies. Together, the sectors create a full disk. It is named for its resemblance to a pie which has been sliced.

To construct it, we use the fact that the total area corresponds to the total number of degrees in the circular arc, namely 360. This 360 is in degree, sometimes written as 360°.

Advantages

1. The pie chart is perhaps the most ubiquitous statistical chart in the business world and the mass media.
2. Pie chart is an effective way of displaying information in some cases, in particular if the intent is to compare the size of a slice with the whole pie, rather than comparing the slices among them.
3. Pie charts work particularly well when the slices represent 25 or 50% of the data.
4. The significance of the information it represents is more easily grasped than by figures it represents.
5. It shows the relationship of the parts to the whole better than the bar chart.

Disadvantages
1. Lack of precise values for individual items.
2. It may be ineffective when there are more than six divisions.
3. Not so easy to draw as a bar chart.
4. It is not so accurate.
5. Values of items cannot be read off from the diagram but must be given.
6. Pie chart is one of the most widely criticized charts, and many statisticians recommend to avoid its use altogether, pointing out in particular that it is difficult to compare different sections of a given pie chart, or to compare data across different pie charts.
7. Statisticians tend to regard pie charts as a poor method of displaying information.
8. It is more difficult for comparisons to be made between the sizes of items in a chart when area is used instead of length.
9. In research performed at AT &T Bell Laboratories, it was shown that comparison by angles was less accurate than comparison with length.
10. Most subjects have difficulty ordering the slices in the pie chart by size; when the bar chart is used the comparison is much easier.

Example
Using the data in the example above on the number of seats won by political parties in Nigeria. Draw a pie chart to represent the data.
Solution

Component Bar Chart
This chart is known as compound bar chart. The chart is made up of bars. Each of these bars represents a class splitted up into constituents’ parts (components). Within each bar, components are always stacked in the same order. A common and helpful arrangement is that of presenting each bar in the order of magnitude with the largest component at the bottom and the smallest at the top. The components are shown with different shades or colours with a proper index.

Example:
The following is a table showing international comparisons of inflation in seven great nations of the world between 1960 – 1972 to 1974:
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>West Germany</td>
<td>3.0</td>
<td>7.0</td>
<td>7.5</td>
</tr>
<tr>
<td>France</td>
<td>4.3</td>
<td>7.1</td>
<td>13.8</td>
</tr>
<tr>
<td>Italy</td>
<td>4.1</td>
<td>10.8</td>
<td>19.3</td>
</tr>
<tr>
<td>Japan</td>
<td>5.5</td>
<td>11.7</td>
<td>25.0</td>
</tr>
<tr>
<td>United States</td>
<td>2.4</td>
<td>5.5</td>
<td>11.5</td>
</tr>
<tr>
<td>Canada</td>
<td>2.3</td>
<td>6.1</td>
<td>10.0</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>3.8</td>
<td>9.1</td>
<td>16.0</td>
</tr>
</tbody>
</table>

Component bar chart of international comparisons of inflations in seven great nations of the world. The series are years.

**Multiple Bar Charts**

This method can be used for data which is made up of two or more components. In this method, the components are shown as separate as separate adjoining bars. The height of each bar represents the actual value of the component. The components are shown by different shades or colours. Where changes in actual values of component figures only are required, multiple bar charts are used.
Example: Use the data above under inflations comparisons.

Multiple bar charts of international comparisons of inflations in seven great nations of the world. The series are years i.e. 1960-72, 1973, and 1974.

Other charts are
1. Z – charts
2. Gantt charts,
3. Semi-logarithmic graphs
4. Ratio scale
5. Lorenz curves
6. Gini – curves
7. Pareto curves

Z – Chart
Z – Chart is a widely used method of showing sales or output data in a form in which it is possible to compare current performance which is subject to seasonal variation with the long term trend of sales or output.

The Z – chart consists of three graphs plotted together on the same axis to form a shape similar to a Z. The three graphs are:

The original data,
The cumulative total and
The moving annual total
When plotted on a graph the three curves form the shape of a ‘Z’. The curve of the original data shows the current fluctuations, the cumulative curve shows the position to date and the trend is indicated by the moving annual total. This chart can be used to compare the basic data with trends of data, such as sales.

**Advantages of Z-chart**

1. Z – chart is capable of showing the performance of a firm/organization to date.
2. The monthly total facilitates a direct check and control of the present performance of an organization.
3. The moving annual total allows us to compare current year performance with the performance at a comparable time in the previous year.

**Disadvantages**

1. Z – chart requires a lot of information to construct.
2. The calculation of the moving annual total which involves the annual totals of the previous year and the current year is tedious and not easy to understand.
3. It has limited applications – it is used mainly.

**Steps in constructing a Z – chart**

Step 1: Collect your data which may be in series of figures on weekly, quarterly or monthly basis as the case may be for two periods/weeks/years.

Step 2: Calculate the annual totals for both the previous and current periods/week/years.

Step 3: Calculate the cumulative total for the current period / weeks / years.

Step 4: Calculate the moving annual total which is the total of output for the 7 days in case of week, 12 months for the year and period. In calculating the moving annual total for say January for the current year (2001), we subtract from the total

Step 5: Choose a convenient scale for the chart and
i. Plot the time series (monthly output/sales) parallel to x-axis.
ii. Plot the cumulative total on the diagonal bar.
iii. Plot the moving annual total on the top bar of the z – chart.

It means that you are going to plot the figures for current year, cumulative total for the current year and moving annual total.

Example:
The output statistics for Virden Plc. Nigeria over a period of two years are as given below.

<table>
<thead>
<tr>
<th>Month</th>
<th>2006 in million</th>
<th>2007 in million</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>56</td>
<td>63</td>
</tr>
<tr>
<td>February</td>
<td>52</td>
<td>54</td>
</tr>
<tr>
<td>March</td>
<td>59</td>
<td>57</td>
</tr>
<tr>
<td>April</td>
<td>61</td>
<td>56</td>
</tr>
<tr>
<td>May</td>
<td>67</td>
<td>63</td>
</tr>
<tr>
<td>June</td>
<td>75</td>
<td>72</td>
</tr>
<tr>
<td>July</td>
<td>85</td>
<td>96</td>
</tr>
<tr>
<td>August</td>
<td>85</td>
<td>96</td>
</tr>
<tr>
<td>September</td>
<td>73</td>
<td>99</td>
</tr>
<tr>
<td>October</td>
<td>62</td>
<td>95</td>
</tr>
<tr>
<td>November</td>
<td>67</td>
<td>83</td>
</tr>
<tr>
<td>December</td>
<td>82</td>
<td>103</td>
</tr>
</tbody>
</table>

Illustrate the progress chart of the output by mans of Z – curve.
### Solution

<table>
<thead>
<tr>
<th>Month</th>
<th>1 2006 (million)</th>
<th>2 2007 (million)</th>
<th>3 Cumulative Total (million)</th>
<th>4 Moving Annual Total (million)</th>
</tr>
</thead>
<tbody>
<tr>
<td>January</td>
<td>56</td>
<td>63</td>
<td>63</td>
<td>831</td>
</tr>
<tr>
<td>February</td>
<td>52</td>
<td>54</td>
<td>117</td>
<td>833</td>
</tr>
<tr>
<td>March</td>
<td>59</td>
<td>57</td>
<td>174</td>
<td>831</td>
</tr>
<tr>
<td>April</td>
<td>61</td>
<td>56</td>
<td>230</td>
<td>826</td>
</tr>
<tr>
<td>May</td>
<td>67</td>
<td>63</td>
<td>293</td>
<td>822</td>
</tr>
<tr>
<td>June</td>
<td>75</td>
<td>72</td>
<td>365</td>
<td>819</td>
</tr>
<tr>
<td>July</td>
<td>85</td>
<td>85</td>
<td>450</td>
<td>819</td>
</tr>
<tr>
<td>August</td>
<td>85</td>
<td>96</td>
<td>546</td>
<td>830</td>
</tr>
<tr>
<td>September</td>
<td>73</td>
<td>99</td>
<td>645</td>
<td>856</td>
</tr>
<tr>
<td>October</td>
<td>62</td>
<td>95</td>
<td>740</td>
<td>889</td>
</tr>
<tr>
<td>November</td>
<td>67</td>
<td>83</td>
<td>823</td>
<td>905</td>
</tr>
<tr>
<td>December</td>
<td>82</td>
<td>103</td>
<td>926</td>
<td>926</td>
</tr>
</tbody>
</table>

824 926

To plot ‘Z’ chart, we need column 2, 3, and 4.

**Key:**
- Series 1 is the actual figures for 2007
- Series 2 is the cumulative total for 2007
- Series 3 is the moving annual total.
Lorenz Curve

The Lorenz curve is a graphical representation of the cumulative distribution function of a probability distribution; it is a graph showing the proportion of the distribution assumed by the bottom \( y \% \) of the values. It is often used to represent income distribution, where it shows for the bottom \( x \% \) of households, what percentage \( y \% \) of the total income they have. The percentage of households is plotted on the \( x \)-axis, the percentage of income on the \( y \)-axis. It can also be used to show distribution of assets. In such use, many economists consider it to be a measure of social inequality. It was developed by Max O. Lorenz in 1905 for representing income distribution.

The Lorenz curve shows the degree of concentration of the measured statistic within the population. The further the curve is away from the diagonal the greater is the degree of concentration. It is a graphical method of showing the deviation from the average of a group of data. The curve is often used to show the level of inequality. For instance, it can show the number of saving against the amount saved. The more equal the distribution of saving, the flatter the curve. If there was equality between the two variables, the curve would be a straight line, equal to the ‘line of equal distribution’. The main application of the Lorenz curve is for comparative purposes. The Lorenz curve gives an immediate impression and it is used for comparison rather than as a quantitative measure of inequality. A coefficient of concentration based on the perpendicular distance from the mid-point of the diagonal to the curve could be calculated for numerical comparison, but so far, no such coefficient has been devised by statisticians and the application of Lorenz curves remains graphical only.

Every point on the Lorenz curve represents a statement like "the bottom \( 20 \% \) of all households have \( 10 \% \) of the total income". A perfectly equal income distribution would be one in which every person has the same income. In this case, the bottom \( N \% \) of society would always have \( N \% \) of the income. This can be depicted by the straight line \( y = x \); called the line of perfect equality.

By contrast, a perfectly unequal distribution would be one in which one person has all the income and everyone else has none. In that case, the curve would be at \( y = 0 \) for all \( x < 100 \% \), and \( y = 100 \% \) when \( x = 100 \% \). This curve is called the line of perfect inequality.
The Gini coefficient is the area between the line of perfect equality and the observed Lorenz curve, as a percentage of the area between the line of perfect equality and the line of perfect inequality. (This equals two times the area between the line of perfect equality and the observed Lorenz curve.) The higher the coefficient, the more unequal the distribution.

Properties
A Lorenz curve always starts at (0,0) and ends at (1,1).

The Lorenz curve is not defined if the mean of the probability distribution is zero or infinite.

The Lorenz curve for a probability distribution is a continuous function. However, Lorenz curves representing discontinuous functions can be constructed as the limit of Lorenz curves of probability distributions, the line of perfect inequality being an example.

If the variable being measured cannot take negative values, the Lorenz curve:
- cannot rise above the line of perfect equality,
- cannot sink below the line of perfect inequality,
- is increasing, and
- is a convex function.

If the variable being measured can take negative values but has a positive mean, then the Lorenz curve will sink below the line of perfect inequality and is a convex function.

If the variable being measured can take negative values and has a negative mean, then the Lorenz curve will be above the line of perfect equality, except at the end points, and is a concave function.

The Lorenz curve is invariant under positive scaling. If $X$ is a random variable, for any positive number $c$ the random variable $cX$ has the same Lorenz curve as $X$.

The Lorenz curve is flipped twice, once about $F = 0.5$ and once about $L = 0.5$, by negation. If $X$ is a random variable with Lorenz curve $L_X(F)$, then $-X$ has the Lorenz curve:

$$L_{-X} = 1 - L_X(1 - F)$$
The Lorenz curve is changed by translations so that the equality gap \( F - L(F) \) changes in proportion to the ratio of the original and translated means. If \( X \) is a random variable with a Lorenz curve \( L_X (F) \) and mean \( \mu_X \), then for any constant \( c \neq -\mu_X \), \( X + c \) has a Lorenz curve defined by:

Like other types of graphs a Lorenz curve is basically used to facilitate comparison of data in income, production, sales etc. with a view to showing the extent of concentration of each phenomenon in the hands of individuals and firms/industries.

It can be used to show the degree of inequality in the distribution of income or wealth of a nation.

It can also be used to show the extent to which the profits of various companies vary.

It can be used to show the extent to which taxes range with income s of the population.

It can also be used to illustrate the variations in the outputs of firms of different sizes.

**Steps in constructing Lorenz curve**

- Compute the cumulative number of establishment and sales value.
- Compute the cumulative percentage sales value.
- Plot the cumulative percentage number of establishments against the cumulative percentage values of sales.
- Draw the line of equal distribution to bring out very clearly the extent of divergence from a uniform distribution.

**Example:**

The table below contains some information on income and the amount of accumulated wealth in a given society.

<table>
<thead>
<tr>
<th>Income in N'000</th>
<th>Number of people</th>
<th>Accumulated wealth N'000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 50</td>
<td>144</td>
<td>32</td>
</tr>
<tr>
<td>50 – 99</td>
<td>54</td>
<td>22</td>
</tr>
<tr>
<td>100 – 149</td>
<td>36</td>
<td>24</td>
</tr>
<tr>
<td>150 – 199</td>
<td>24</td>
<td>20</td>
</tr>
<tr>
<td>200 – 249</td>
<td>18</td>
<td>24</td>
</tr>
<tr>
<td>250 – 299</td>
<td>15</td>
<td>26</td>
</tr>
<tr>
<td>300 – 349</td>
<td>9</td>
<td>52</td>
</tr>
</tbody>
</table>

Represent the above data by means of a Lorenz Curve.
Solution

<table>
<thead>
<tr>
<th>Income in (N'000)</th>
<th>No of People (f)</th>
<th>Accumulated Wealth (N'000)</th>
<th>Cumulative Wealth (N'000)</th>
<th>Cumulative Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 50</td>
<td>144</td>
<td>3</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>50 – 99</td>
<td>54</td>
<td>22</td>
<td>54</td>
<td>27</td>
</tr>
<tr>
<td>100 – 149</td>
<td>36</td>
<td>24</td>
<td>78</td>
<td>39</td>
</tr>
<tr>
<td>150 – 199</td>
<td>24</td>
<td>20</td>
<td>98</td>
<td>49</td>
</tr>
<tr>
<td>200 – 249</td>
<td>18</td>
<td>24</td>
<td>122</td>
<td>61</td>
</tr>
<tr>
<td>250 – 299</td>
<td>15</td>
<td>26</td>
<td>148</td>
<td>74</td>
</tr>
<tr>
<td>300 - 349</td>
<td>9</td>
<td>52</td>
<td>200</td>
<td>100</td>
</tr>
</tbody>
</table>

% of wealth

The curve above indicates that the distribution of wealth is not very equal. If there was ‘equality’ then 10% of the income earners would have 10% of the total wealth and 50% of the earners would have 50% of the total wealth and so on.

Semi Logarithmic Graph

When comparing two time series, it is often instructive to determine the proportional rather than the actual rate of increase in the number of comprising the series. The rate of increase of a variable is measured per unit time. In order to show the proportional rate of a variable graphically, the logarithm of the variable (usually the dependent variable) is measured along one of the axes of the graph rather than its actual magnitude. The other axis of the graph is usually a natural scale along which the actual magnitudes of the other variable a (usually the independent variable) are
measured. Such a graph is known as a semi-logarithmic graph, one of the axes being logarithm scale whilst the other is a natural scale.

The semi-logarithmic graph (or semi-log or ratio scale graph) is used to show the rate of change in data, rather than changes in actual amounts (which are shown on natural scale graphs). Usually only one axis (the y or vertical axis) is measured in a logarithmic or ratio scale. Therefore, the graph is called a semi-log graph. The important factor on a semi-log graph is the degree of slope of the curve. The curve of the usual graph measures the magnitude at any point while the log graph shows at any point the percentage change from the last point. The distinguishing feature of a semi-logarithmic graph is that a variable which is increasing at a constant proportional rate is represented by a straight-line on the graph; although on the natural scale graph, the same variable will be represented by a curve which is not straight, but concave upwards, showing that the actual rate of increase of the variable is increasing.

A semi-logarithmic graph can be constructed either by finding the logarithm of each number in the series and measuring these logarithms along the vertical axis of the graph so that the equal distances represent equal increments in the logarithm; or, alternatively, the numbers may be plotted directly onto specially ruled semi-logarithmic graph paper. The advantage of the former method is that if the number are all closed together in magnitude, the relation between them can be brought out clearly by choosing a large scale to represent the logarithms. Specially prepared graph paper, on the other hand, does not provide a choice of scale, though this is no disadvantage if the numbers are well spaced out in magnitude.
Example:

The table below shows the production and export of cocoa in Osun state of Nigeria from 1999 to 2008.

<table>
<thead>
<tr>
<th>Year</th>
<th>Production in tonnes</th>
<th>Export in Tonnes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Logarithm</td>
</tr>
<tr>
<td>1999</td>
<td>708</td>
<td>2.85</td>
</tr>
<tr>
<td>2000</td>
<td>861</td>
<td>2.94</td>
</tr>
<tr>
<td>2001</td>
<td>1052</td>
<td>3.02</td>
</tr>
<tr>
<td>2002</td>
<td>1190</td>
<td>3.08</td>
</tr>
<tr>
<td>2003</td>
<td>1353</td>
<td>3.13</td>
</tr>
<tr>
<td>2004</td>
<td>1004</td>
<td>3.00</td>
</tr>
<tr>
<td>2005</td>
<td>1249</td>
<td>3.10</td>
</tr>
<tr>
<td>2006</td>
<td>1608</td>
<td>3.21</td>
</tr>
<tr>
<td>2007</td>
<td>1868</td>
<td>3.27</td>
</tr>
</tbody>
</table>

The graph of natural figures for cocoa production and exports for the year 1999 through to 2007.

**Gini Coefficient**

The Gini coefficient is a measure of statistical dispersion most prominently used as a measure of inequality of income distribution or inequality of wealth distribution. It is defined as a ratio with values between 0 and 1: A low Gini coefficient indicates more equal income or wealth distribution, while a high Gini coefficient indicates more unequal distribution. 0 corresponds to perfect equality (everyone having exactly the same income) and 1 corresponds to perfect inequality (where one person has all the income, while everyone else has zero income). The Gini coefficient requires that no one have a negative net income or wealth. Worldwide, Gini coefficients range from approximately 0.232 in Denmark to 0.707 in Namibia.

The Gini index is the Gini coefficient expressed as a percentage, thus Denmark's Gini index is 23.2% (Mathematically, this is equal to the Gini coefficient of 0.232, but the percentage sign is often omitted in the Gini index.)

The Gini coefficient was developed by the Italian statistician Corrado Gini and published in his 1912 paper "Variability and Mutability" (Italian: *Variabilità e mutabilità*).

The Gini coefficient is also commonly used for the measurement of the discriminatory power of rating systems in credit risk management. Since gini coefficient addresses wealth inequality it may be important to understand what a transformative asset is. Transformative assets increase the gini coefficient as they provide a family or individual with a wealth advantage over most persons.

**Calculation**

The Gini coefficient is defined as a ratio of the areas on the Lorenz curve diagram. If the area between the line of perfect equality and Lorenz curve is $A$, and the area under the Lorenz curve is $B$, then the Gini coefficient is $\frac{A}{(A+B)}$. Since $A+B = 0.5$, the Gini coefficient, $G = \frac{A}{0.5} = 2A = 1-2B$. If the Lorenz curve is represented by the function $Y = L(X)$, the value of $B$ can be found with integration and:

$$
G = 1 - 2 \int_0^1 L(X) dX
$$
In some cases, this equation can be applied to calculate the Gini coefficient without direct reference to the Lorenz curve. For example:

For a population uniform on the values \( y_i, i = 1 \) to \( n \), indexed in non-decreasing order (\( y_i \leq y_{i+1} \)):

\[
G = \frac{1}{n} \left[ \frac{n + 1}{2} \sum_{i=1}^{n} (n + 1 - i)y_i - \frac{\sum_{i=1}^{n} y_i}{n} \right]
\]

- For a cumulative distribution function \( F(y) \) that is piecewise differentiable, has a mean \( \mu \), and is zero for all negative values of \( y \):
- Since the Gini coefficient is half the relative mean difference, it can also be calculated using formulas for the relative mean difference. For a random sample \( S \) consisting of values \( y_i, i = 1 \) to \( n \), that are indexed in non-decreasing order (\( y_i \leq y_{i+1} \)), the statistic:

\[
G(S) = \frac{1}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (y_i + y_j)
\]

is a consistent estimator of the population Gini coefficient, but is not, in general, unbiased. Like, \( G \), \( G(S) \) has a simpler form:

There does not exist a sample statistic that is in general an unbiased estimator of the population Gini coefficient, like the relative mean difference.

Sometimes the entire Lorenz curve is not known, and only values at certain intervals are given. In that case, the Gini coefficient can be approximated by using various techniques for interpolating the missing values of the Lorenz curve. If \((X_k, Y_k)\) are the known points on the Lorenz curve, with the \(X_k\) indexed in increasing order (\( X_{k-1} < X_k \)), so that:

- \( X_k \) is the cumulated proportion of the population variable, for \( k = 0, \ldots, n \), with \( X_0 = 0, X_n = 1 \).
- \( Y_k \) is the cumulated proportion of the income variable, for \( k = 0, \ldots, n \), with \( Y_0 = 0, Y_n = 1 \).
If the Lorenz curve is approximated on each interval as a line between consecutive points, then the area B can be approximated with trapezoids and:

is the resulting approximation for G. More accurate results can be obtained using other methods to approximate the area B, such as approximating the Lorenz curve with a quadratic function across pairs of intervals, or building an appropriately smooth approximation to the underlying distribution function that matches the known data. If the population means and boundary values for each interval are also known, these can also often be used to improve the accuracy of the approximation.

The Gini coefficient calculated from a sample is a statistic and its standard error, or confidence intervals for the population Gini coefficient, should be reported. These can be calculated using bootstrap techniques but those proposed have been mathematically complicated and computationally onerous even in an era of fast computers. Ogwang (2000) made the process more efficient by setting up a “trick regression model” in which the incomes in the sample are ranked with the lowest income being allocated rank 1. The model then expresses the rank (dependent variable) as the sum of a constant A and a normal error term whose variance is inversely proportional to \( y_i \).

Ogwang showed that G can be expressed as a function of the weighted least squares estimate of the constant A and that this can be used to speed up the calculation of the jackknife estimate for the standard error. Giles (2004) argued that the standard error of the estimate of A can be used to derive that of the estimate of G directly without using a jackknife at all. However it has since been argued that this is dependent on the model’s assumptions about the error distributions (Ogwang 2004) and the independence of error terms (Reza & Gastwirth 2006) and that these assumptions are often not valid for real data sets. It may therefore be better to stick with jackknife methods such as those proposed by Yitzhaki (1991) and Karagiannis and Kovacevic (2000). The debate continues.

The gini coefficient can be calculated if you know the mean of a distribution, the number of people (or percentiles), and the income of each person (or percentile). Princeton development economist Angus Deaton (1997, 139) has simplified the Gini calculation to one easy formula:

where \( u \) is mean income of the population, \( P_i \) is the income rank \( P \) of person \( i \), with income \( X \), such that the richest person receives a rank of 1
and the poorest a rank of N. This effectively gives higher weight to poorer people in the income distribution, which allows the Gini to meet the Transfer Principle

**Income Gini indices in the world**

A complete listing is in list of countries by income equality; the article economic inequality discusses the social and policy aspects of income and asset inequality.

While most developed European nations tend to have Gini indices between 24 and 36, the United States' and Mexico's Gini indices are both above 40, indicating that the United States and Mexico have greater inequality. Using the Gini can help quantify differences in welfare and compensation policies and philosophies. However it should be borne in mind that the Gini coefficient can be misleading when used to make political comparisons between large and small countries.

The Gini index for the entire world has been estimated by various parties to be between 56 and 66.

**Correlation with per-capita GDP**

Poor countries (those with low per-capita GDP) generally have higher Gini indices, spread between 40 and 65, with extremes at 25 and 71, while rich countries generally have lower Gini indices (under 40). The lowest Gini coefficients (under 30) can be found in continental Europe. Overall, there is a clear negative correlation between Gini coefficient and GDP per capita, the U.S.A, Hongkong and Singapore being rich exceptions with high Gini coefficients.

In many of the former socialist countries and in-development capitalist countries (e.g., Brazil), the sizeable underground economy may hide income for many. In such a case, earning/wealth statistics over-represent certain income ranges (i.e., in lower-income regions), and may alter the Gini coefficient either up or down.

**US income Gini indices over time**

Gini indices for the United States at various times, according to the US Census Bureau:

1967: 39.7 (first year reported)
1968: 38.6 (lowest index reported)
Advantages of Gini Coefficient as a Measure of Inequality
The Gini coefficient's main advantage is that it is a measure of inequality by means of a ratio analysis, rather than a variable unrepresentative of most of the population, such as per capita income or gross domestic product.

It can be used to compare income distributions across different population sectors as well as countries, for example the Gini coefficient for urban areas differs from that of rural-areas in many countries (though the United States' urban and rural Gini coefficients are nearly identical).

It is sufficiently simple that it can be compared across countries and be easily interpreted. GDP statistics are often criticised as they do not represent changes for the whole population; the Gini coefficient demonstrates how income has changed for poor and rich. If the Gini coefficient is rising as well as GDP, poverty may not be improving for the majority of the population.

The Gini coefficient can be used to indicate how the distribution of income has changed within a country over a period of time, thus it is possible to see if inequality is increasing or decreasing.

The Gini Coefficient Satisfies Four Important Principles:
*Anonymity*: it does not matter who the high and low earners are.

*Scale independence*: the Gini coefficient does not consider the size of the economy, the way it is measured, or whether it is a rich or poor country on average.

*Population independence*: it does not matter how large the population of the country is.
Transfer principle: if income (less than the difference), is transferred from a rich person to a poor person the resulting distribution is more equal.

Disadvantages of Gini Coefficient as a Measure of Inequality
The Gini coefficient of different sets of people cannot be averaged to obtain the Gini coefficient of all the people in the sets: if a Gini coefficient were to be calculated for each person it would always be zero. For a large, economically diverse country, a much higher coefficient will be calculated for the country as a whole than will be calculated for each of its regions. (The coefficient is usually applied to measurable nominal income rather than local purchasing power, tending to increase the calculated coefficient across larger areas.)

For this reason, the scores calculated for individual countries within the EU are difficult to compare with the score of the entire US: the overall value for the EU should be used in that case, which is still much lower than the United States'. Using decomposable inequality measures (e.g. the Theil index $T$ converted by $1 - e^{-T}$ into a inequality coefficient) averts such problems.

The Lorenz curve may understate the actual amount of inequality if richer households are able to use income more efficiently than lower income households. From another point of view, measured inequality may be the result of more or less efficient use of household incomes.

Economies with similar incomes and Gini coefficients can still have very different income distributions. This is because the Lorenz curves can have different shapes and yet still yield the same Gini coefficient.

It measures current income rather than lifetime income. A society in which everyone earned the same over a lifetime would appear unequal because of people at different stages in their life; a society in which students study rather than save can never have a coefficient of 0. However, Gini coefficient can also be calculated for any kind of distribution, e.g. for wealth.

Problems in using the Gini coefficient
Gini coefficients do include income gained from wealth; however, the Gini coefficient is used to measure net income more than net worth, which can be misinterpreted. For example, Sweden has a low Gini coefficient for income distribution and a higher Gini coefficient for wealth (the wealth
inequality is low by international standards, but still significant: 5% of Swedish household shareholders hold 77% of the share value owned by households). In other words, the Gini income coefficient should not be interpreted as measuring effective egalitarianism. Distribution of stock ownership does not appear to correlate to many recognized indicators of egalitarianism.

Too often only the Gini coefficient is quoted without describing the proportions of the quantiles used for measurement. As with other inequality coefficients, the Gini coefficient is influenced by the granularity of the measurements. For example, five 20% quantiles (low granularity) will usually yield a lower Gini coefficient than twenty 5% quantiles (high granularity) taken from the same distribution. This is an often encountered problem with measurements.

Care should be taken in using the Gini coefficient as a measure of egalitarianism, as it is properly a measure of income dispersion. Two equally egalitarian countries with different immigration policies may have different Gini coefficients.

**General Problems of Measurement**

Comparing income distributions among countries may be difficult because benefits systems may differ. For example, some countries give benefits in the form of money while others give food stamps, which might not be counted by some economists and researchers as income in the Lorenz curve and therefore not taken into account in the Gini coefficient. U.S. counts income before benefits, while France counts it after benefits, making US appear more unequal vis-a-vis France than it is.

The measure will give different results when applied to individuals instead of households. When different populations are not measured with consistent definitions, comparison is not meaningful.

As for all statistics, there may be systematic and random errors in the data. The meaning of the Gini coefficient decreases as the data become less accurate. Also, countries may collect data differently, making it difficult to compare statistics between countries.

As one result of this criticism, in addition to or in competition with the Gini coefficient, *entropy measures* are frequently used (e.g. the Theil Index and the index of Atkinson). These measures attempt to compare the distribution of resources by intelligent agents in the market with a
maximum entropy random distribution, which would occur if these agents acted like non-intelligent particles in a closed system following the laws of statistical physics.

**Post-Test**

The following is the list of Nigeria Cities that are over a million and their populations:

<table>
<thead>
<tr>
<th>City</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lagos</td>
<td>7,937,932</td>
</tr>
<tr>
<td>Kano</td>
<td>3,848,885</td>
</tr>
<tr>
<td>Ibadan</td>
<td>3,078,400</td>
</tr>
<tr>
<td>Kaduna</td>
<td>1,652,844</td>
</tr>
<tr>
<td>Port-Harcourt</td>
<td>1,320,214</td>
</tr>
<tr>
<td>Benin City</td>
<td>1,051,600</td>
</tr>
<tr>
<td>Maiduguri</td>
<td>1,044,497</td>
</tr>
<tr>
<td>Zaria</td>
<td>1,018,827</td>
</tr>
</tbody>
</table>

Represent the data using Bar chart and Pie chart

**References**


LECTURE SIX

Measures of Central Tendency

Introduction
In the previous lecture, we have studied how to collect raw data, its classification and tabulation in a useful form, which contributes in solving many problems of statistical concern. Yet, this is not sufficient, for in practical purposes, there is need for further condensation, particularly when we want to compare two or more different distributions. We may reduce the entire distribution to one number which represents the distribution. This is called an average.

Objectives
At the end of this lecture, readers are expected to have mastered the following:
1. how to calculate arithmetic mean using direct and assumed mean;
2. computations of median and mode;
3. calculation of harmonic mean, weighted mean and geometric mean; and
4. computations of all these measures from both the grouped and ungrouped data.

Pre-Test
1. In a test marked out of 10, a group of pupils obtained the following marks:
   3,4,6,3,4,3,5,6,7,8,9,5,9,10,7,8,2,6,5,4,10,5,6,7,3,8,9,4,2,6.
   i. Prepare a frequency table without grouping into classes.
ii. Find the mean, mode and median of the marks.
iii. Find the number of people who scored at least 5 marks.

2. The table shows the distribution of wages in a garment factory in Lagos.

<table>
<thead>
<tr>
<th>Wages (N’000)</th>
<th>No of workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 – 20</td>
<td>6</td>
</tr>
<tr>
<td>21 – 40</td>
<td>14</td>
</tr>
<tr>
<td>41 – 60</td>
<td>9</td>
</tr>
<tr>
<td>61 – 80</td>
<td>6</td>
</tr>
<tr>
<td>81 - 100</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>40</strong></td>
</tr>
</tbody>
</table>

i. Calculate the mean monthly wages in the factory
ii. Use the table to draw a histogram, and hence or otherwise, estimate the mode and the median from the graph.

3. The monthly forecast of temperatures (in °C) in fifty selected countries in African sub region are as follows:

12.5  3.4  11.6  11.1  18.4  22.8  14  4.8  24  14.8  19
13.5  21.5  13.9  14.7  18.7  3  8.4  19.9  6.6  12.6  7.9
14.6  17.6  23.9  12  13.4  12.4  26.7  8.1  4.2  9.4  29
11.5  25  27.8  17.7  15  28.9  6  13.5  20  30  14.3
21.2  13.9  9.6  12.9  16.5  13

a. Construct a frequency distribution for the data by using class intervals of 1 – 5, 6 – 10, etc.
b. Use your table in (a) to calculate:
   i. the mean monthly temperature using an assumed mean of 13°C.
   ii. the modal monthly temperature
   iii. Use your results in (i) and (ii) to guess the shape of the distribution graphically.
The Meaning of Measures of Central Tendency

Measures of central tendency are also called a measure of locations. This is also referred to as average. A single value which can be considered as typical or representative of a set of observations and around which the observations can be considered as centered is called an average. It is a value which is typical or representative of a set of data. Since such typical value tends to lie centrally within a set of data arranged according to magnitude.

Several types of averages can be defined, the most fundamental measures of tendencies are:

i. Arithmetic mean  
ii. Median  
iii. Mode  
iv. Geometric mean  
v. Harmonic mean  
vi. Weighted averages

However the most common measures of locations are Arithmetic mean, median and mode. We therefore, consider the Arithmetic mean.

**Arithmetic Mean**

Arithmetic mean is the most commonly used average. It can be defined as the sum of aggregate of a series of items divided by their number. Thus, you are expected to add all observations (values of all items) together and divide this sum by the number of observations (or items). For instance, the mean of a set of numbers \( x_1, x_2, x_3, \ldots, x_n \) is denoted by \( \bar{x} \) and is defined as

\[
\bar{x} = \frac{x_1 + x_2 + x_3 + \ldots + x_n}{n}
\]

Example; Find the average mean of 2, 4, 5, 7, 9, 5, 3

\[
\bar{x} = \frac{\sum x}{n} = \frac{2 + 4 + 5 + 7 + 9 + 5 + 3}{7} = \frac{35}{7} = 5
\]
This is an example of ungrouped form of data. In an ungrouped data, it is sometimes too large and the interpretation thereof may be meaningless but the actual values are used in calculations.

If the numbers $x_1, x_2, x_3, \ldots, x_n$ occur $f_1, f_2, f_3, \ldots, f_n$ times respectively (i.e. occur with frequencies $f_1, f_2, f_3, \ldots, f_n$), the arithmetic mean is:

$$X = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \ldots + f_nx_n}{f_1 + f_2 + f_3 + \ldots + f_n} = \frac{\sum fx}{\sum f}$$

**Example:**
The following are the ages in years of 47 workers in a private organization:

<table>
<thead>
<tr>
<th>Ages</th>
<th>No of workers</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>12</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>31</td>
<td>6</td>
</tr>
<tr>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
</tr>
<tr>
<td>34</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
</tr>
</tbody>
</table>

Calculate the mean age

**Solution:** Using direct method

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>fx</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>6</td>
<td>168</td>
</tr>
<tr>
<td>29</td>
<td>12</td>
<td>348</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>210</td>
</tr>
<tr>
<td>31</td>
<td>6</td>
<td>186</td>
</tr>
<tr>
<td>32</td>
<td>7</td>
<td>224</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>99</td>
</tr>
<tr>
<td>34</td>
<td>6</td>
<td>204</td>
</tr>
<tr>
<td>Total</td>
<td>47</td>
<td>1439</td>
</tr>
</tbody>
</table>
Mean age = \( \frac{\sum fx}{\sum f} = \frac{1439}{47} = 30.62 \approx 31 \text{ years} \)

**Indirect Method**

Given that an assumed mean is used, we shall have the formula

\[
\text{Mean} = A + \frac{\sum fd}{\sum f}
\]

In the example given above, suppose an assumed mean of 30 years is chosen, then the table will look like this:

<table>
<thead>
<tr>
<th>X</th>
<th>d = X - 30</th>
<th>f</th>
<th>fd</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>-2</td>
<td>6</td>
<td>-12</td>
</tr>
<tr>
<td>29</td>
<td>-1</td>
<td>12</td>
<td>-12</td>
</tr>
<tr>
<td>30</td>
<td>0</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>31</td>
<td>1</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>32</td>
<td>2</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td>34</td>
<td>4</td>
<td>6</td>
<td>24</td>
</tr>
</tbody>
</table>

\[
\text{Mean} = A + \frac{\sum fd}{\sum f} = 30 + \frac{29}{47} \approx 30.6170 \approx 31 \text{ years.}
\]

If a grouped data is given like the one below
**Example:** The table below gives the distribution of marks scored by fifty students in a school examination.

<table>
<thead>
<tr>
<th>Scores</th>
<th>Number of candidates</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 and under 20</td>
<td>2</td>
</tr>
<tr>
<td>20 and under 30</td>
<td>2</td>
</tr>
<tr>
<td>30 and under 40</td>
<td>3</td>
</tr>
<tr>
<td>40 and under 50</td>
<td>4</td>
</tr>
<tr>
<td>50 and under 60</td>
<td>6</td>
</tr>
<tr>
<td>60 and under 70</td>
<td>12</td>
</tr>
<tr>
<td>70 and under 80</td>
<td>14</td>
</tr>
<tr>
<td>80 and under 90</td>
<td>4</td>
</tr>
<tr>
<td>90 and under 100</td>
<td>3</td>
</tr>
</tbody>
</table>

Calculate the mean mark using

i. Direct method and

ii. Indirect method

**Solution**

<table>
<thead>
<tr>
<th>Scores</th>
<th>Class mark (x)</th>
<th>Frequency (f)</th>
<th>Fx</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 and under 20</td>
<td>15</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>20 and under 30</td>
<td>25</td>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>30 and under 40</td>
<td>35</td>
<td>3</td>
<td>105</td>
</tr>
<tr>
<td>40 and under 50</td>
<td>45</td>
<td>4</td>
<td>180</td>
</tr>
<tr>
<td>50 and under 60</td>
<td>55</td>
<td>6</td>
<td>330</td>
</tr>
<tr>
<td>60 and under 70</td>
<td>65</td>
<td>12</td>
<td>780</td>
</tr>
<tr>
<td>70 and under 80</td>
<td>75</td>
<td>14</td>
<td>1050</td>
</tr>
<tr>
<td>80 and under 90</td>
<td>85</td>
<td>4</td>
<td>340</td>
</tr>
<tr>
<td>90 and under 100</td>
<td>95</td>
<td>3</td>
<td>285</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>50</td>
<td></td>
<td>3150</td>
</tr>
</tbody>
</table>
i. **Direct method**

The mean = \( \frac{\sum fx}{\sum f} = 3150 \div 50 = 63 \)

ii. **Indirect method**

Let the assumed mean = 45

<table>
<thead>
<tr>
<th>Class mark x</th>
<th>d = x - 45</th>
<th>f</th>
<th>fd</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td>-30</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>25</td>
<td>-20</td>
<td>2</td>
<td>-40</td>
</tr>
<tr>
<td>35</td>
<td>-10</td>
<td>3</td>
<td>30</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>55</td>
<td>10</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>65</td>
<td>20</td>
<td>12</td>
<td>240</td>
</tr>
<tr>
<td>75</td>
<td>30</td>
<td>14</td>
<td>420</td>
</tr>
<tr>
<td>85</td>
<td>40</td>
<td>4</td>
<td>160</td>
</tr>
<tr>
<td>95</td>
<td>50</td>
<td>3</td>
<td>150</td>
</tr>
</tbody>
</table>

The mean = \( A + \frac{\sum fd}{\sum f} = 45 + \frac{900}{50} = 45 + 18 = 63 \)

The results obtained from the two methods are the same; hence, you are free to use any of the methods when the need arises.

The following are the advantages and disadvantages of mean

**Advantages**

1. It is a true representation of all the items in the data.
2. It is widely understood and the basic calculation is straightforward
3. It makes use of all the data in the group and it can be determined with mathematical precision.
4. It is very useful in further statistical work.
Disadvantages
1. It cannot be obtained graphically
2. It is affected by extreme values
3. It may not correspond to an actual value and this may make it appear unrealistic.

Properties of Arithmetic Mean
1. The sum of the deviations, of all the values of x, from their arithmetic mean is zero. Justification: \[ \sum f(x-x) = \sum fx - x \sum f = 0. \] since x is constant, \[ x = \frac{\sum fx}{\sum f} \] therefore \[ \sum fx = x \sum f \]
2. The product of the arithmetic mean and the number of items gives the total of all items \[ x = \frac{\sum fx}{\sum f} = x = \sum fx = x \sum f \]
3. If \( m_1 \) and \( x_2 \) are the arithmetic mean of two samples of sizes \( n_1 \) and \( n_2 \) respectively then, arithmetic mean \( M \) of the distribution combining the two can be calculated as

\[ M = \frac{n_1 m_1 + n_2 m_2}{n_1 + n_2} \]

this is called the combined mean of the two samples.

Example: The average age of three groups of students having 25, 30 and 32 students respectively are 60, 50 and 30. Find the average age of all the 140 students taken together.

Solution:
Let \( M \) be the average age of all 140 students taken together

<table>
<thead>
<tr>
<th>Group 1</th>
<th>Group 2</th>
<th>Group 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( m_1 )</td>
<td>( m_2 )</td>
<td>( m_3 )</td>
</tr>
<tr>
<td>25</td>
<td>30</td>
<td>32</td>
</tr>
</tbody>
</table>

No. of students \( n_1 = 60 \), \( n_2 = 50 \), \( n_3 = 30 \)

\[ M = \frac{n_1 m_1 + n_2 m_2 + n_3 m_3}{n_1 + n_2 + n_3} = \frac{60 \times 25 + 50 \times 30 + 30 \times 32}{60 + 50 + 30} = \frac{1500 + 1500 + 960}{140} = \frac{3960}{140} = 28.2857 \]
\[ \frac{1500 + 1500 + 960}{140} = \frac{3900}{140} = 27.86 \approx 28 \text{ years} \]

**Example:** The mean of a certain number of observation is 40. If two or more items with values 50 and 64 are added to this data, the mean rises to 42. Find the number of items in the original data.

**Solution**

Let \( n \) be the number of observations whose mean \( m = 40 \)

\[
m = \frac{\sum x_i}{n} \quad \Rightarrow \quad \sum x_i = nm = n(40)
\]

= 40\( n \) total of \( n \) values

Two more items of values 50 and 64 are added therefore, total of \( (n+2) \) values:

\[
= \sum x_i + 50 + 64
\]

\[
= 40n + 50 + 64
\]

\[
= 40n + 114
\]

Now new mean is 42

\[
\therefore \text{New } m = \text{new total of } (n+2) \text{ values}
\]

\[
\therefore 42 = \frac{40n + 114}{n + 2}
\]

\[
\therefore 42n + 84 = 40n + 114
\]

\[
\therefore 2n = 30
\]

\[
\therefore n = 15
\]

Therefore, the number of items in the original data is 15

114
**Median**

The median of a collection of data is the middlemost measurement when the data are arranged according to size. It conveys the notion of the middle value, dividing the distribution into two halves. Median is used in many contexts such as median annual salary, median family income, median age of University graduates, and so on.

When there is an odd number of observations in the data, the median of a set of values is defined as the middle value when the observations are arranged in an array in order of magnitude from the smallest to the largest (or vice versa). If the number of observation is even, then the median is defined as the mean of the two middle values.

**Example:** The set of numbers

\[
\begin{align*}
4 & \quad 6 & \quad 7 & \quad 8 & \quad 9 \\
\end{align*}
\]

has 7 at the middle, therefore, 7 is the median.

Also, in the set of numbers

\[
\begin{align*}
2 & \quad 3 & \quad 5 & \quad 7 & \quad 8 & \quad 9 \\
\end{align*}
\]

has 5 and 7 at the middle, therefore the median is

\[
\frac{5 + 7}{2} = 6
\]

has its median.

But if the distribution is grouped thus

<table>
<thead>
<tr>
<th>x</th>
<th>f</th>
<th>cf</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>22</td>
</tr>
</tbody>
</table>

The median is \((N/2)\)th item = \(\sum f/2 = 22/2 = 11\)th item which is the item that falls on cumulative frequency 15. The corresponding value is 3, hence, if the values are arranged, 3 will be at the middle.

If we have frequency distributions like the one below. The median is found using the formula:

\[
\text{Median} = L_{me} + \left\{\frac{N/2 - C_{fb}}{f_{m}}\right\}c
\]

where \(L_{me}\) is the lower class boundary of the median class (i.e. the class containing the median class).
\(N\) is the number of items in the distribution (i.e. total frequency)
\(C_{fb}\) is the cumulative frequency below the median class
\(f_{m}\) is the frequency of the median class.
\(C\) is the size of the median class (i.e. the class interval of the median class).

Geometrically, the median is the value of \(X\) (abscissa) corresponding to the vertical line which divides a histogram into two parts having equal areas.

**Example:** the table below gives the distribution of salary been received monthly by staff of a corporation in Lagos.

<table>
<thead>
<tr>
<th>Salary in (N'000)</th>
<th>No of workers (f)</th>
<th>Cumulative frequency</th>
<th>Class boundary</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 24</td>
<td>14</td>
<td>14</td>
<td>19.5 – 24.5</td>
</tr>
<tr>
<td>25 – 29</td>
<td>18</td>
<td>32</td>
<td>24.5 – 29.5</td>
</tr>
<tr>
<td>30 – 34</td>
<td>11</td>
<td>43</td>
<td>29.5 – 34.5</td>
</tr>
<tr>
<td>35 – 39</td>
<td>25</td>
<td>68</td>
<td>34.5 – 39.5</td>
</tr>
<tr>
<td>40 – 44</td>
<td>9</td>
<td>77</td>
<td>39.5 – 44.5</td>
</tr>
<tr>
<td>45 – 49</td>
<td>3</td>
<td>80</td>
<td>44.5 – 49.5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>80</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The problem now is how to locate the median class. This is done by dividing the sum-total of frequency by 2, that is,

\[ N/2 = 80/2 = 40^{th} \text{ item} \]

This falls on the third class which is between 29.5 and 34.5. The lower class boundary of the median class is 29.5, therefore, the median is calculated thus

\[
\text{Median} = 29.5 + \left( \frac{40 - 32}{11} \right) \times 5
\]

\[
= 29.5 + 8 \times 5
\]

\[
= 33.14 (\text{'000})
\]

That is the median salary is ₦33,140.00 in that organization. This divides the distribution in half, so that of the 80 workers, half earn less than ₦33,140.00 and half will earn more. In this sense, the median provides a good representation of grouped frequency distribution.

Median can also be estimated from cumulative frequency curve (ogive).

**Advantages of the Median**

1. It is not influenced by extreme values.
2. It is straightforward to calculate even if not all values are known or when there are irregular class intervals.
3. It is often an actual value and even when it is not it may look representative and realistic.
4. It can be obtained graphically from a frequency curve called ogive.

**Disadvantages**

1. It is not used in further statistical work.
2. It represents a value which may not be characteristic of the group.
3. It cannot be used to determine the value of all the items, the number of items multiplied by the median will not give the total for the data, therefore it is not suitable for further arithmetic calculations.

**Mode**

The mode can be defined as the most frequently occurring value in a distribution. It is the value which occurs most often. For continuous distribution, it is the value of variable corresponding to the highest point or peak of the frequency curve. The mode may not exist in some distribution and if it does, it may not be unique.

**Example:**

<table>
<thead>
<tr>
<th>The sizes of packing shirts</th>
<th>Number of shirts produced</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>6</td>
</tr>
<tr>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>16</td>
<td>10</td>
</tr>
<tr>
<td>161/2</td>
<td>7</td>
</tr>
<tr>
<td>17</td>
<td>5</td>
</tr>
<tr>
<td>18</td>
<td>3</td>
</tr>
</tbody>
</table>

The mode in this example is 16, because it has highest frequency of 10.

But for the grouped data, mode is calculated using the formula below:

\[
\text{Mode} = L_{mo} + \left( \frac{f_1 - f_0}{2f_1 - f_0 - f_2} \right) \times c
\]

Or

\[
\text{Mode} = L_{mo} + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c
\]

Where

- \(f_0\) is the frequency of the group before the modal class.
- \(f_1\) is the frequency of the modal class.
\( f_2 \) is the frequency of the class after the modal class.
\( C \) is the class size.

**Example:** Recall the table in the example above

\[
\text{Mode} = 34.5 + \frac{[25 – 11] \times 5}{[50 – 11 – 9]}
\]
\[
= 34.5 + \frac{14}{6}
\]
\[
= 34.5 + 2.33
\]
\[
= 36.83 (\text{'000})
\]
\[
= \text{₦36,830.00} \text{ is the mode and it is the modal salary.}
\]

The above data can be used to draw a histogram in order to estimate the mode from it.

<table>
<thead>
<tr>
<th>Class boundary</th>
<th>No of workers (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>19.5 – 24.5</td>
<td>14</td>
</tr>
<tr>
<td>24.5 – 29.5</td>
<td>18</td>
</tr>
<tr>
<td>29.5 – 34.5</td>
<td>11</td>
</tr>
<tr>
<td>34.5 – 39.5</td>
<td>25</td>
</tr>
<tr>
<td>39.5 – 44.5</td>
<td>9</td>
</tr>
<tr>
<td>44.5 – 49.5</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>80</td>
</tr>
</tbody>
</table>
Advantages of Mode

1. It is a commonly used average
2. It has practical uses, for instance, cars, shirts, shoes are made to modal sizes.
3. The mode is often an actual value and therefore may appear to be realistic and sensible.
4. Modal information can often be supplied quickly by people who have experience in a particular area.
5. It is very easy to find for a set of ungrouped data.
6. It is not affected by extreme values.

Figure: Histogram estimating modal salary
7. it is the only measure of central tendency that can be used with nominal data.

**Disadvantages**

1. It is not useful for further statistical work.
2. It may not be well defined and can often be a matter of judgment.
3. It does not include all the values in the distribution.
4. There may be more than one mode.
5. The mode is greatly subject to sample fluctuations and is therefore not recommended to be used as the only measure of central tendency.

In a normal distribution, the mean, median, and mode are identical

**Geometric Mean \([G]\)**

The geometric mean \(G\) of a set of \(n\) number \(x_1, x_2, \ldots, x_n\) is the \(n\)th root of the product of the numbers

\[
G = \sqrt[n]{x_1, x_2, \ldots, x_n}
\]

Or \(G = [x_1, x_2, \ldots, x_n]^{1/n}\)

geometric mean is the \(n\)th root of the product of the scores. Thus, the geometric mean of the scores: 1, 2, 3, and 10 is the fourth root of 1 x 2 x 3 x 10 which is the fourth root of 60 which equals 2.78. The formula can be written as:

Geometric mean = \((\prod X)^{1/n}\)

where \(\Pi X\) means to take the product of all the values of \(X\). The geometric mean can also be computed by:

1. taking the logarithm of each number
2. computing the arithmetic mean of the logarithms raising the base used to take the logarithms to the arithmetic mean
**Example:** Find the geometric mean of the following numbers: 2, 4, 8

\[
G = (2 \times 4 \times 8)^{1/3} = (64)^{1/3} = 4.
\]

Naturally, you get the same result using logs base 10 as shown below.

<table>
<thead>
<tr>
<th>X</th>
<th>Log(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0000</td>
</tr>
<tr>
<td>2</td>
<td>0.30103</td>
</tr>
<tr>
<td>3</td>
<td>0.47712</td>
</tr>
<tr>
<td>10</td>
<td>1.00000</td>
</tr>
</tbody>
</table>

Geometric mean = 2.78

Arithmetic mean = 0.44454. \(10^{0.44454} = 2.78\)

If any one of the scores is zero then the geometric mean is zero. The geometric mean does not make sense if any scores are less than zero.

The geometric mean is less affected by extreme values than is the arithmetic mean and is useful as a measure of central tendency for some positively skewed distributions.

The geometric mean is an appropriate measure to use for averaging rates. For example, consider a stock portfolio that began with a value of N1,000 and had annual returns of 13%, 22%, 12%, -5%, and -13%. The table below shows the value after each of the five years.

<table>
<thead>
<tr>
<th>Year</th>
<th>Return</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>13%</td>
<td>1,130</td>
</tr>
<tr>
<td>2</td>
<td>22%</td>
<td>1,379</td>
</tr>
<tr>
<td>3</td>
<td>12%</td>
<td>1,544</td>
</tr>
<tr>
<td>4</td>
<td>-5%</td>
<td>1,467</td>
</tr>
<tr>
<td>5</td>
<td>-13%</td>
<td>1,276</td>
</tr>
</tbody>
</table>
The question is how to compute annual rate of return? The answer is to compute the geometric mean of the returns. Instead of using the percents, each return is represented as a multiplier indicating how much higher the value is after the year. This multiplier is 1.13 for a 13% return and 0.95 for a 5% loss. The multipliers for this example are 1.13, 1.22, 1.12, 0.95, and 0.87. The geometric mean of these multipliers is 1.05. Therefore, the average annual rate of return is 5%. The following table shows how a portfolio gaining 5% a year would end up with the same value (N1,276) as the one shown above.

<table>
<thead>
<tr>
<th>Year</th>
<th>Return</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5%</td>
<td>1,050</td>
</tr>
<tr>
<td>2</td>
<td>5%</td>
<td>1,103</td>
</tr>
<tr>
<td>3</td>
<td>5%</td>
<td>1,158</td>
</tr>
<tr>
<td>4</td>
<td>5%</td>
<td>1,216</td>
</tr>
<tr>
<td>5</td>
<td>5%</td>
<td>1,276</td>
</tr>
</tbody>
</table>

**Harmonic Mean**

The harmonic mean is used to take the mean of sample sizes. If there are k samples each of size n, then the harmonic mean is defined as:

$$n_h = \frac{k}{\frac{1}{n_1} + \frac{1}{n_2} + \ldots + \frac{1}{n_k}}$$

For the numbers 1, 2, 3, and 10, the harmonic mean is:

$$n_h = \frac{4}{\frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{10}} = 2.069.$$  

= 2.069. This is less than the geometric mean of 2.78 and the arithmetic mean of 4.
Post-Test

1. A firm of Auditors recorded the following distribution of auditing time of its Clients’ account.

<table>
<thead>
<tr>
<th>Auditing time in hours (x)</th>
<th>Number of accounts (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 – 19</td>
<td>2</td>
</tr>
<tr>
<td>20 – 29</td>
<td>3</td>
</tr>
<tr>
<td>30 – 39</td>
<td>5</td>
</tr>
<tr>
<td>40 – 49</td>
<td>8</td>
</tr>
<tr>
<td>50 – 59</td>
<td>10</td>
</tr>
<tr>
<td>60 – 69</td>
<td>11</td>
</tr>
<tr>
<td>70 - 79</td>
<td>13</td>
</tr>
</tbody>
</table>

a. Using the deviate \( d = (x - 45.5)/10 \) determine, to the nearest hour, the mean and mean deviation of the distribution.

b. Obtain to the nearest half hour, the median and the other quartiles of the distribution and, hence, determine the quartile measure of skewness.

2. The average weekly domestic expenditures of 130 housewives selected at random in Ibadan metropolis are stated as follows:

<table>
<thead>
<tr>
<th>Expenditures in N’000</th>
<th>Number of housewives</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 – 9</td>
<td>15</td>
</tr>
<tr>
<td>10 – 14</td>
<td>18</td>
</tr>
<tr>
<td>15 – 19</td>
<td>17</td>
</tr>
<tr>
<td>20 – 24</td>
<td>19</td>
</tr>
<tr>
<td>25 – 29</td>
<td>12</td>
</tr>
<tr>
<td>30 – 34</td>
<td>14</td>
</tr>
<tr>
<td>35 – 39</td>
<td>12</td>
</tr>
<tr>
<td>40 – 49</td>
<td>11</td>
</tr>
<tr>
<td>50 – 59</td>
<td>8</td>
</tr>
<tr>
<td>60 - 69</td>
<td>4</td>
</tr>
</tbody>
</table>

a. Compute the mean, median and the mode of the distribution

b. What are the merits and demerits of these measures.

c. Compute the relative frequency and cumulative frequency of this distribution and estimate the proportion of housewives whose weekly domestic expenditures are at most N34,000.
3. A new generation bank classified the income of its borrowers as follows:

<table>
<thead>
<tr>
<th>Annual income in N’000</th>
<th>Number of borrowers</th>
</tr>
</thead>
<tbody>
<tr>
<td>15 but less than 20</td>
<td>10</td>
</tr>
<tr>
<td>20 but less than 25</td>
<td>12</td>
</tr>
<tr>
<td>25 but less than 30</td>
<td>23</td>
</tr>
<tr>
<td>30 but less than 35</td>
<td>35</td>
</tr>
<tr>
<td>35 but less than 40</td>
<td>45</td>
</tr>
<tr>
<td>40 but less than 45</td>
<td>30</td>
</tr>
<tr>
<td>45 but less than 50</td>
<td>17</td>
</tr>
<tr>
<td>50 but less than 55</td>
<td>8</td>
</tr>
</tbody>
</table>

a. Using an assumed mean of 32.5 and a scale factor of 5, determine
   i. The arithmetic mean
   ii. The median
   iii. The total income of these borrowers per annum
   iv. How many borrowers has this bank?

b. What are the merits of arithmetic mean?

4. A sample survey recently carried out amongst manufacturing companies in Nigeria showed the following profit made that particular year.

<table>
<thead>
<tr>
<th>Profits in N’000</th>
<th>Number of companies</th>
</tr>
</thead>
<tbody>
<tr>
<td>500 – 600</td>
<td>8</td>
</tr>
<tr>
<td>600 – 700</td>
<td>12</td>
</tr>
<tr>
<td>700 – 800</td>
<td>14</td>
</tr>
<tr>
<td>800 – 900</td>
<td>7</td>
</tr>
<tr>
<td>900 – 1000</td>
<td>10</td>
</tr>
<tr>
<td>1000 – 1100</td>
<td>9</td>
</tr>
<tr>
<td>1100 – 1200</td>
<td>11</td>
</tr>
<tr>
<td>1200 – 1300</td>
<td>9</td>
</tr>
<tr>
<td>1300 – 1400</td>
<td>15</td>
</tr>
<tr>
<td>1400 - 1500</td>
<td>5</td>
</tr>
</tbody>
</table>
1. Calculate the sample mean profit and median profit.
2. Compute the total profits of the company.

Summary
Of the five measures of central tendency discussed, the mean is by far the most widely used. It takes every score into account, is the most efficient measure of central tendency for normal distributions and is mathematically tractable making it possible for statisticians to develop statistical procedures for drawing inferences about means. On the other hand, the mean is not appropriate for highly skewed distributions and is less efficient than other measures of central tendency when extreme scores are possible. The geometric mean is a viable alternative if all the scores are positive and the distribution has a positive skew. The median is useful because its meaning is clear and it is more efficient than the mean in highly-skewed distributions. However, it ignores many scores and is generally less efficient than the mean. The mode can be informative but should almost never be used as the only measure of central tendency since it is highly susceptible to sampling fluctuations.

Reference


LECTURE SEVEN

Measures of Dispersion

Introduction
Simpson and Kaika said, “An average alone does not tell the full story. It is hardly fully representative of a mass, unless we know the manner in which the individual item scatter around it….a further description of a series is necessary, if we are to gauge how representative the average is.” What we are saying here is that the measures of central tendencies (i.e. means) indicate the general magnitude of the data and locate only the center of a distribution of measures. They do not establish the degree of variability or the spread out or scatter of the individual items and their deviation from the means. This is the focus of our discussion here. Before this let us run through the objectives of the lecture.

Objectives
At the completion of this lecture, you should be able to:

1. discuss the meaning of the measures of dispersion;
2. discuss the methods of computing dispersion;
3. explain the concepts of range and related terms;
4. compute quartiles; and
5. calculate coefficient of variation and its usefulness.

Pre-Test
1. What is dispersion? Enumerate its measurements.
2. Define the following terms:
   a. Range
   b. Deviation
c. Coefficient of range
d. Coefficient of variation
e. Skewness

3. Following are the ages of selected Psychologists randomly selected from various Universities in the South-West:

44  49  35  41  46  25  47  60  54  46  33  38

Find the
a. The range
b. Mean age
c. Mean deviation
d. Standard deviation and
e. Coefficient of variation of the distribution

CONTENT
What is Dispersion?
The degree to which numerical data tend to spread about an average value is called the variation or dispersion of the data.

The arithmetic mean of the deviations of the values of the individual items from the measure of a particular central tendency used. Thus the dispersion is also known as the “average of the second degree”.

The word dispersion may also be used to indicate the spread of the data. The basic property of dispersion is that “the value which indicates the extent to which all the values are dispersed about the central value in a particular distribution is called dispersion, or variation or scatter or deviation.” In measuring dispersion, it is imperative to know the amount of variation (absolute measure) and the degree of variation (relative measure). In the former case we consider the range, mean deviation, variance, standard deviation etc. In the latter case we consider the coefficients of range, mean deviation, variation etc.

Methods of Dispersion
To study dispersion, two methods are outlined below:

a. Mathematical methods
b. Graphical methods
a. **Mathematical methods**: In this method, we can study the degree and extent of dispersion. In this category, common measures of dispersion are:
   i. Range
   ii. Quartile deviation
   iii. Mean deviation
   iv. Variance and standard variation
   v. Coefficient of variation

b. **Graphical methods**: Here, we study only the extent of dispersion, whether it is more or less, but the degree of dispersion is not possible. In this case we resort to Lorenz curve or cumulative percentage curve.

Some other authors divided methods of computing dispersion into methods of limits and methods of averages.

Methods of limits are the range, interquartile range and percentile range.

Method of averages are quartile deviation, mean deviation, standard deviation and other measures. Let us study these measures one after the other.

1. **Range**

   In any statistical series, the difference between the largest and the smallest values is called the range. The range is the simplest measure of spread or dispersion. The range can be a useful measure of spread because it is so easily understood. However, it is very sensitive to extreme scores since it is based on only two values. The range should almost never be used as the only measure of spread, but can be informative if used as a supplement to other measures of spread such as the standard deviation or semi-interquartile range.

   \[ \text{L} = \text{Largest value of the series} \]

   Thus Range \( (R) = \text{L} - \text{S} \) \{ 

   \[ \text{S} = \text{Smallest value of the series} \]
**Coefficient of Range:** The relative measure of the range. It is used in the comparative study of the dispersion coefficient of

\[
\text{Range} = \frac{L - S}{L + S}
\]

The range is efficient when \( n \leq 10 \), otherwise it is not good as it ignores all the values in between. It is commonly used in statistical quality control

**Example:** Find the range and the coefficient of the range of the following items:

\[
2 \quad 3 \quad 3 \quad 5 \quad 6 \quad 7 \quad 8 \quad 10 \quad 12
\]

**Solution**

Range = Highest – Lowest = 12 – 2 = 10

Coefficient of the range = \( \frac{L - S}{L + S} = \frac{12 - 2}{12 + 2} = \frac{10}{14} = 0.7143 \)

**Mean Deviation**

Mean absolute deviation is the average amount of variations (scatter) of the items in a distribution from either the mean or the median or the mode, ignoring the signs of these deviations.

It is calculated thus

\[
MD = \frac{\sum |x - m|}{N}
\]

or \( \frac{MD = \frac{\sum f |x - m|}{\sum f}}{\sum f} \)

**Example:** The following figures give the price of roast chicken (in naira) at Caristico supermarket, Ibadan during six weeks.

\[
1200 \quad 1150 \quad 1350 \quad 1550 \quad 1530 \quad 1620
\]

Find the mean deviation of the distribution.
Solution:
Mean = \( \frac{\sum x}{n} = \frac{8350}{6} = 1400 \)

\[ : \text{MD} = \frac{\sum |x - m|}{n} = \frac{|1200 - 1400| + |1150 - 1400| + |1350 - 1400| + |1550 - 1400| + |1530 - 1400| + |1620 - 1400|}{6} \]

= \frac{200 + 250 + 150 + 130 + 220}{6} = \frac{1000}{6} = 166.67

Standard Deviation and Variance
The variance and the closely-related standard deviation are measures of how spread out a distribution is. In other words, they are measures of variability.

The concept of variance is of great importance in advanced work where it is possible to split the total into several parts, each attributable to one of the factors causing variations in their original series.

The variance is computed as the average squared deviation of each number from its mean. For example, for the numbers 1, 2, and 3, the mean is 2 and the variance is:

\[ \sigma^2 = \frac{(1-2)^2 + (2-2)^2 + (3-2)^2}{3} = 0.667 \]

The formula (in summation notation) for the variance in a population is

\[ \sigma^2 = \frac{\sum (X - \mu)^2}{N} \]

where \( \mu \) is the mean and \( N \) is the number of scores.

When the variance is computed in a sample, the statistic

\[ s^2 = \frac{\sum (X - M)^2}{N} \]
(where $M$ is the mean of the sample) can be used. $S^2$ is a biased estimate of $\sigma^2$, however. By far the most common formula for computing variance in a sample is:

$$s^2 = \frac{\sum (x - M)^2}{N-1}$$

which gives an unbiased estimate of $\sigma^2$. Since samples are usually used to estimate parameters, $s^2$ is the most commonly used measure of variance. Calculating the variance is an important part of many statistical applications and analyses. It is the first step in calculating the standard deviation.

**Standard Deviation**

The standard deviation formula is very simple: it is the square root of the variance. It is the most commonly used measure of spread.

$$\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{\sum f}} \quad \text{or} \quad \sigma = \sqrt{\frac{\sum f x^2 - (\sum f x)^2}{\sum f}}$$

An important attribute of the standard deviation as a measure of spread is that if the mean and standard deviation of a normal distribution are known, it is possible to compute the percentile rank associated with any given score. In a normal distribution, about 68% of the scores are within one standard deviation of the mean and about 95% of the scores are within two standard deviations of the mean.

The standard deviation has proven to be an extremely useful measure of spread in part because it is mathematically tractable. Many formulas in inferential statistics use the standard deviation.

Although less sensitive to extreme scores than the range, the standard deviation is more sensitive than the semi-interquartile range. Thus, the standard deviation should be supplemented by the semi-interquartile range when the possibility of extreme scores is present.

If variable $Y$ is a linear transformation of $X$ such that:

$$Y = bX + A,$$

then the variance of $Y$ is:

$$\sigma_y^2 = b^2 \sigma_x^2$$
where \( \sigma^2 \) is the variance of \( X \).

The standard deviation of \( Y \) is \( b\sigma_x \) where \( \sigma_x \) is the standard deviation of \( X \).

**Standard Deviation as a Measure of Risk**

The standard deviation is often used by investors to measure the risk of a stock or a stock portfolio. The basic idea is that the standard deviation is a measure of volatility: the more a stock's returns vary from the stock's average return, the more volatile the stock. Consider the following two stock portfolios and their respective returns (in per cent) over the last six months. Both portfolios end up increasing in value from \( N1,000 \) to \( N1,058 \). However, they clearly differ in volatility. Portfolio A's monthly returns range from -1.5% to 3% whereas Portfolio B's range from -9% to 12%. The standard deviation of the returns is a better measure of volatility than the range because it takes all the values into account. The standard deviation of the six returns for Portfolio A is 1.52; for Portfolio B it is 7.24.

**Merits of standard deviation**

i. It is rigidly defined and based on all observations.

ii. It is amenable to further algebraic treatment.

iii. It is not affected by sampling fluctuations

iv. It is less erratic.

**Demerits**

i. It is difficult to understand

ii. It gives greater weight to extreme values.

**Coefficient of Variation (CV)**

To compare the variations (dispersion) of two different series, relative measures of standard deviation must be calculated. This is known as coefficient of variation or the coefficient of standard deviation. Its formula is

\[
\sigma / \mu \times 100
\]

Thus it is defined as the ratio standard deviation to its mean.
It is given as a percentage and is used to compare the consistency or variability of two more series. The higher the coefficient of variation, the higher the variability and lower the coefficient of variation (CV), the higher is the consistency of the data.

**Quartiles, Deciles and Percentiles**

If a set of data is arranged in order of magnitude, the middle value (or arithmetic mean of the two middle values) which divides the set into two equal parts is the median. By extending this idea we can think of those values which divide the set into four equal parts. These values denote Q1, Q2 and Q3 are called the first, second and third quartiles respectively, the value Q2 being equal to the median.

Similarly, the values which divide the data into ten equal parts are called deciles and are denoted by D1, D2, ……. D10, while the values that divide the data into hundred equal parts are called percentiles and are denoted P1, P2, ……. P99. The 5th decile and 50th percentile correspond to the median. The 25th and 75th percentiles correspond to the first and the third quartiles respectively.

**Example:** The wages distribution [₹'000] of 39 employees of a banking institutions are as follows:

<table>
<thead>
<tr>
<th>31</th>
<th>33</th>
<th>28</th>
<th>31</th>
<th>35</th>
<th>34</th>
<th>29</th>
<th>28</th>
<th>29</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>32</td>
<td>28</td>
<td>31</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>33</td>
<td>29</td>
<td>29</td>
</tr>
<tr>
<td>31</td>
<td>32</td>
<td>35</td>
<td>35</td>
<td>30</td>
<td>32</td>
<td>30</td>
<td>28</td>
<td>28</td>
<td>34</td>
</tr>
<tr>
<td>34</td>
<td>32</td>
<td>28</td>
<td>31</td>
<td>33</td>
<td>33</td>
<td>34</td>
<td>35</td>
<td>30</td>
<td></td>
</tr>
</tbody>
</table>

Compute the following from the distribution

i. The upper quartile
ii. The lower quartile
iii. The sixth decile
iv. The 40th percentile
v. The quartile deviation
vi. Interpercentile range
vii. Semi-interquartile range
Solution

<table>
<thead>
<tr>
<th>Wages (₦'000)</th>
<th>Tally</th>
<th>No of workers (f)</th>
<th>Cumulative frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>28</td>
<td>///</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>29</td>
<td>///</td>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>30</td>
<td>///</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>31</td>
<td>///</td>
<td>6</td>
<td>19</td>
</tr>
<tr>
<td>32</td>
<td>///</td>
<td>5</td>
<td>24</td>
</tr>
<tr>
<td>33</td>
<td>///</td>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>34</td>
<td>///</td>
<td>5</td>
<td>34</td>
</tr>
<tr>
<td>35</td>
<td>///</td>
<td>5</td>
<td>39</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>39</strong></td>
<td></td>
</tr>
</tbody>
</table>

i. Upper quartile = \( \frac{3}{4} \) of 39 = 29.25 this correspond to ₦34,000

ii. Lower quartile = \( \frac{1}{4} \) of 39 = 9.75 this correspond to ₦29,000

iii. The sixth decile = \( \frac{6}{10} \) of 39 = 23.4\(^{th}\) item which correspond to ₦32,000.

iv. The 40\(^{th}\) percentile = \( \frac{40}{100} \) of 39 = 15.6\(^{th}\) item which correspond to ₦31,000.

v. The quartile deviation = upper quartile – lower quartile = 34000 – 29000 = ₦5000

vi. Interpercentile range = \( Q_9 - Q_1 \) = 35000 – 28000 = ₦7000.00

vii. Semi-interquartile range = \( \frac{Q_3 - Q_1}{2} \) = \( \frac{34000 - 29000}{2} \) = 2500 = ₦2,500

Example: Referring to the table of salary of workers in an establishment. See the table below:

From the table, calculate the mean absolute deviation, variance, standard deviation and coefficient of variation.
<table>
<thead>
<tr>
<th>Salary in (N’000)</th>
<th>No of workers (f)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 24</td>
<td>14</td>
</tr>
<tr>
<td>25 – 29</td>
<td>18</td>
</tr>
<tr>
<td>30 – 34</td>
<td>11</td>
</tr>
<tr>
<td>35 – 39</td>
<td>25</td>
</tr>
<tr>
<td>40 – 44</td>
<td>9</td>
</tr>
<tr>
<td>45 – 49</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>80</strong></td>
</tr>
</tbody>
</table>

Solution:

<table>
<thead>
<tr>
<th>Salary (N’000)</th>
<th>No of Workers (f)</th>
<th>Class Marks (x)</th>
<th>fx</th>
<th>x - µ</th>
<th>x - µ ²</th>
<th>(x - µ)²</th>
<th>f(x - µ)²</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 – 24</td>
<td>4</td>
<td>22</td>
<td>88</td>
<td>-12.6</td>
<td>158.76</td>
<td>50.4</td>
<td>158.76</td>
</tr>
<tr>
<td>25 – 29</td>
<td>8</td>
<td>27</td>
<td>216</td>
<td>-7.6</td>
<td>57.76</td>
<td>60.4</td>
<td>57.76</td>
</tr>
<tr>
<td>30 – 34</td>
<td>11</td>
<td>32</td>
<td>352</td>
<td>-2.6</td>
<td>2.6</td>
<td>28.6</td>
<td>36</td>
</tr>
<tr>
<td>35 – 39</td>
<td>15</td>
<td>37</td>
<td>555</td>
<td>2.4</td>
<td>5.76</td>
<td>36</td>
<td>5.76</td>
</tr>
<tr>
<td>40 – 44</td>
<td>9</td>
<td>42</td>
<td>378</td>
<td>7.4</td>
<td>54.76</td>
<td>66.6</td>
<td>54.76</td>
</tr>
<tr>
<td>45 – 49</td>
<td>3</td>
<td>47</td>
<td>141</td>
<td>12.4</td>
<td>153.76</td>
<td>37.2</td>
<td>153.76</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>50</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td><strong>279.2</strong></td>
<td></td>
</tr>
</tbody>
</table>

Mean \( \mu = \frac{\sum fx}{\sum f} = \frac{1730}{50} = 34.6 \) or N’34,600.00

Mean Absolute Deviation = \( \frac{\sum f |x - \mu|}{\sum f} = \frac{279.2}{50} = 5.584 \) or N’5584.00
\[ \sigma^2 = \frac{\sum f(x - \mu)^2}{\sum f} = \frac{2212}{50} = 44.24 = \frac{\sum f}{50} \]

\[ \sigma = \sqrt{44.24} = 6.6513 \]

Coefficient of variation = \( \sigma / \mu \times 100 = 6.6513/34.6 = 19.22\% \)

**Combined standard deviation:** If two sets containing \( n_1 \) and \( n_2 \) items having means \( x_1 \) and \( x_2 \) and standard deviations \( \sigma_1 \) and \( \sigma_2 \) respectively are taken together, then

1. Mean of the combined data is \( \bar{x} = \frac{n_1x_1 + n_2x_2}{n_1 + n_2} \)
2. Standard deviation of the combined data is \( \sigma = \sqrt{\frac{n_1(\sigma_1^2 + d_1^2) + n_2(\sigma_2^2 + d_2^2)}{n_1 + n_2}} \)

where \( d_1 = x_1 - \bar{x} \) and \( d_2 = x_2 - \bar{x} \)

**Skew**

A distribution is skewed if one of its tails is longer than the other. The first distribution shown has a positive skew. This means that it has a long tail in the positive direction. The distribution below it has a negative skew since it has a long tail in the negative direction. Finally, the third distribution is symmetric and has no skew. Distributions with positive skew are sometimes called "skewed to the right" whereas distributions with negative skew are called "skewed to the left."
Distributions with positive skew are more common than distributions with negative skews. One example is the distribution of income. Most people make under $40,000 a year, but some make quite a bit more with a small number making many millions of dollars per year. The positive tail therefore extends out quite a long way whereas the negative tail stops at zero.

For a more psychological example, a distribution with a positive skew typically results if the time it takes to make a response is measured. The longest response times are usually much longer than typical response times whereas the shortest response times are seldom much less than the typical response time. A histogram of the author's performance on a perceptual motor task in which the goal is to move the mouse to and click on a small target as quickly as possible is shown below. The X axis shows times in milliseconds.
Negatively skewed distributions do occur, however. Consider the following frequency polygon of test grades on a statistics test where most students did very well but a few did poorly. It has a large negative skew.

![Frequency Polygon](image)

**Kurtosis**

Skew can be calculated as:

\[
\text{skew} = \frac{\sum (X-\mu)^3}{N\sigma^3}
\]

where \(\mu\) is the mean and \(\sigma\) is the standard deviation.

The normal distribution has a skew of 0 since it is a symmetric distribution.

As a general rule, the mean is larger than the median in positively skewed distributions and less than the median in negatively skewed distributions. There are counter examples. For example it is not uncommon for the median to be higher than the mean in a positively skewed bimodal distribution or with discrete distributions.

Kurtosis is based on the size of a distribution's tails. Distributions with relatively large tails are called "leptokurtic"; those with small tails are called "platykurtic." A distribution with the same kurtosis as the normal distribution is called "mesokurtic."

The following formula can be used to calculate kurtosis:

\[
\text{kurtosis} = \frac{\sum (X-\mu)^4}{N\sigma^4} - 3
\]

where \(\sigma\) is the standard deviation. The kurtosis of a normal distribution is 0.
The following two distributions have the same variance, approximately the same skew, but differ markedly in kurtosis.

Summary
The standard deviation is by far the most widely used measure of spread. It takes every score into account, has extremely useful properties when used with a normal distribution, and is tractable mathematically and, therefore; it appears in many formulas in inferential statistics. The standard deviation is not a good measure of spread in highly-skewed distributions and should be supplemented in those cases by the semi-interquartile range.

The range is a useful statistic, but it cannot stand alone as a measure of spread since it takes into account only two scores.

The semi-interquartile range is rarely used as a measure of spread, in part because it is not very mathematically tractable. However, it is influenced less by extreme scores than the standard deviation, is less subject to sampling fluctuations in highly-skewed distributions, and has a good intuitive meaning. It should be used to supplement the standard deviation in most cases.

Post-Test
1. The Accountant at a Bank put the debit situation in the bank as follows:

<table>
<thead>
<tr>
<th>Debit Balance in N’000</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 – 19</td>
<td>11</td>
</tr>
<tr>
<td>20 – 29</td>
<td>13</td>
</tr>
<tr>
<td>30 – 39</td>
<td>6</td>
</tr>
<tr>
<td>40 – 49</td>
<td>5</td>
</tr>
<tr>
<td>50 – 59</td>
<td>3</td>
</tr>
<tr>
<td>60 – 69</td>
<td>2</td>
</tr>
</tbody>
</table>
a. Compute:
   i. The mean
   ii. The standard deviation
b. Compute relevant indices of:
   i. Skewness and
   ii. Relative variation

2. The distribution of remuneration per annum of regular employees of a Nigerian company earning above N64,000.00 is given as follows:

<table>
<thead>
<tr>
<th>Remuneration in N'000</th>
<th>Number of Employees</th>
</tr>
</thead>
<tbody>
<tr>
<td>60 – 70</td>
<td>19</td>
</tr>
<tr>
<td>70 – 80</td>
<td>17</td>
</tr>
<tr>
<td>80 – 90</td>
<td>8</td>
</tr>
<tr>
<td>90 – 100</td>
<td>13</td>
</tr>
<tr>
<td>100 – 110</td>
<td>11</td>
</tr>
<tr>
<td>110 – 120</td>
<td>10</td>
</tr>
<tr>
<td>120 – 130</td>
<td>4</td>
</tr>
<tr>
<td>130 - 140</td>
<td>2</td>
</tr>
<tr>
<td>140 – 150</td>
<td>3</td>
</tr>
<tr>
<td>150 – 160</td>
<td>2</td>
</tr>
<tr>
<td>160 – 170</td>
<td>4</td>
</tr>
<tr>
<td>170 – 180</td>
<td>3</td>
</tr>
<tr>
<td>180 – 190</td>
<td>1</td>
</tr>
<tr>
<td>190 – 200</td>
<td>1</td>
</tr>
</tbody>
</table>

a. Determine the following:
   i. Total remuneration
   ii. Average remuneration
   iii. Median remuneration
   iv. The standard deviation of the distribution
   v. The coefficient of variation of the distribution.
b. A cooperative society classified the income of its borrowers as follows:

<table>
<thead>
<tr>
<th>Annual income in N'000</th>
<th>Number of borrowers</th>
</tr>
</thead>
<tbody>
<tr>
<td>25 but less than 30</td>
<td>7</td>
</tr>
<tr>
<td>30 but less than 35</td>
<td>20</td>
</tr>
<tr>
<td>35 but less than 40</td>
<td>36</td>
</tr>
<tr>
<td>40 but less than 45</td>
<td>24</td>
</tr>
<tr>
<td>45 but less than 50</td>
<td>21</td>
</tr>
<tr>
<td>50 but less than 55</td>
<td>7</td>
</tr>
<tr>
<td>55 but less than 60</td>
<td>5</td>
</tr>
</tbody>
</table>

a. How many people borrowed from this cooperative society?

b. Using an assumed mean of 37.5 and a scale factor of 5, determine:
   i. The arithmetic mean
   ii. The standard deviation
   iii. The median of income distribution.
   iv. Determine from these measures, the skewness of the distribution.

c. Draw an ogive of the distribution and determine graphically the upper quartile. What is the interpretation of this quartile?

3. Give one advantage and one disadvantage of the standard deviation as a measure of the spread of a set of data values.

The maximum temperatures in degrees Celsius at eight places in Nigeria on a August day in 2008 were as follows:

20 13 18 17 21 23 15 19

a. Calculate the mean and standard deviation of these temperatures

b. What is the range and mid-range of the distribution?

c. Determine the distribution coefficient of variation.

4. Below is the average age of ten politicians randomly selected from a community in Oyo state:

54 59 35 41 46 25 47 60 54 46
Find the
i. The range
ii. Mean age
iii. Mean deviation
iv. Standard deviation and
v. Coefficient of variation of the distribution.

References


LECTURE EIGHT

Measures of Relationships

Introduction
This life is a matter of interaction of many factors all aims at achieving a common goal. The most important problems of interest to social scientists is that of finding the relationship between variables. In this lecture, we want to see how variables relate, how changes in one variable affect another or rather to say is there any relationship between variables. This may as well predict the value of one variable from some functional relationship. Also, this lecture will examine the degree of relationship between two variables. Let us see the highlight of objectives in this lecture.

Objectives
At the end of this lecture, you should be able to:
1. define and describe regression;
2. define and describe correlation;
3. compute regression coefficient;
4. derive a regression equation;
5. compute correlation coefficients and their interpretations; and
6. plot scatter plot to know the type of relationship that exists between variables.

Pre-Test
1. Distinguish between regression coefficient and correlation coefficient.
2. Discuss briefly the features of a correlation coefficient.
3. A random of 10 Economics teacher was selected to know whether there is a relationship between their income and their propensity to save (pps).

<table>
<thead>
<tr>
<th>Income x in N’000</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>35</th>
<th>40</th>
<th>45</th>
<th>50</th>
<th>60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Propensity to save y</td>
<td>0.3</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.6</td>
<td>0.2</td>
<td>0.5</td>
<td>0.4</td>
<td>0.3</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Use the method of least squares to determine an equation for the best straight line for these data. i.e. Y = a + bx.

4. Explain the term regression line. Make use of the line to predict the pps for an income of:
   i. N55,000
   ii. N70,000.

5. The following table gives the advertisement spending and the volume of sales generated for a particular product.

<table>
<thead>
<tr>
<th>Amount spent on Adverts (N’000) x</th>
<th>Values of sales made (N’000) y</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>30</td>
</tr>
<tr>
<td>45</td>
<td>50</td>
</tr>
<tr>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>65</td>
<td>70</td>
</tr>
<tr>
<td>80</td>
<td>80</td>
</tr>
<tr>
<td>85</td>
<td>90</td>
</tr>
<tr>
<td>90</td>
<td>100</td>
</tr>
<tr>
<td>100</td>
<td>110</td>
</tr>
<tr>
<td>75</td>
<td>100</td>
</tr>
<tr>
<td>60</td>
<td>70</td>
</tr>
<tr>
<td>35</td>
<td>50</td>
</tr>
<tr>
<td>20</td>
<td>35</td>
</tr>
</tbody>
</table>

   a. Display the data on a scatter diagram and comment on the features.
   b. Determine the equation of the regression line of the form Y = α + βx, α and β are constants.

146
c. Use the equation to estimate for the sales made if N=50,000 is expended on advertisement.

d. Calculate the correlation coefficient as well as coefficient of determination.

CONTENT
Regression
Regression attempts to show relationship between two variables by providing a mean line which best indicates the trend of the points or coordinates on a graph. The aim is to minimize the total divergence of the points or coordinates from the line. The main purpose of regression is to fit a functional relationship connecting the dependent variable and the independent variable and thereafter use it to predict values for the dependent variable for known values of the independent variable. Mathematically, it has been found that the best line is one that minimizes the total of the squared deviations. This is known as the method of least squares.

Regression Analysis
Regression analysis is a statistical tool which helps to predict one variable from the other variable or variables on the basis of the assumed nature of the relationship between them. Regression analysis is the mathematical process of using observations to find the line of best fit through the data in order to make estimates and predictions about the behaviour of the variables. This line of best fit may be linear (straight) or curvilinear to some mathematical formula. The variable being predicted is known as the dependent variable while the other variable(s) is/are referred to as the independent variable(s).

Curve Fitting/Scatter Diagram
To aid in determining an equation connecting variables, the first step is the collection of data showing corresponding values of the variables under consideration. The next step is to plot the points \((x_1, y_1), (x_2, y_2), (x_3, y_3), \ldots, (x_n, y_n)\) on a rectangular coordinates system. The resulting set of points is sometimes called a scatter diagram or scatter plot. Scatter diagram is a graph which shows the relationship between two variables.
From the scatter diagram, it is often possible to visualize a smooth curve approximating the data.

Scatter plots should almost always be constructed when the relationship between two variables is of interest. Statistical summaries are no substitute for a full plot of the data.

We have the following curves showing relationship of the variables.

![Scatter plots]

a. Positive linear relationship  
b. Negative linear relationship

c. No relationship  
d. Curvilinear relationship

**Linear Regression**

Linear regression is a form of regression analysis in which observational data are modeled by a function which is a linear combination of the model parameters and depends on one or more independent variables. In simple linear regression the model function represents a straight line. The results of data fitting are subject to statistical analysis.
The Regression Equation/Model

This is the equation which gives the relationship between variables when there is not a ‘unique’ relationship between them. The data consists of $n$ values $y_1, \ldots, y_n$ taken from observations of the dependent variable (response variable) $y$. The dependent variable is subject to error. This error is assumed to be random variable, with a mean zero. Systematic error may be present but its treatment is outside the scope of regression analysis. The independent variable (explanatory variable) $x$, is error-free. If this is not so, modeling should be done using errors-in-variables model techniques. The independent variables are also called regressors, exogenous variables, input variables and predictor variables. In simple linear regression the data model is written as:

$$ Y = \alpha + \beta x + \epsilon_i, \quad i = 1, 2, \ldots, n $$

Where
- $Y$ is the dependent variable
- $x$ is the independent variable
- $\alpha$ is the intercept (note that this can be represented with a or any letter of the alphabets.
- $\beta$ is the slope of the line. It is sometimes called regression coefficient and it can also be represented by $b$ or any letter of the alphabets.
- $\epsilon$ is the error that is independently and normally distributed with mean 0 and constant variance $\sigma^2$.

Above equation can be written as

$$ Y = a + bx. $$

This is an example of a simple bivariate linear regression model. It is also called estimated model for sample data. There are also multiple and non-linear models.

Least Squares Estimate of Regression Model

This method of fitting a line uses the values of all observations. The basic idea of estimating quantities by minimizing a sum of squares of deviations
(errors) was first proposed by Gauss, and is called the principle of least squares.

We assume that the particular values $x_1y_1$ of two variables $X,Y$ are related by the equation

$$ Y = \beta x + \epsilon_i $$

In which $\beta$ is the slope of the underlying straight line relationship and $\epsilon_i$ is the error of the determination of $y_i$. We have that

$$ E(y_i) = E(\beta x_i + \epsilon_i) $$
$$ = E(\beta x_i) + E(\epsilon_i) $$
$$ = \beta x_i \text{ since } E(\epsilon_i) = 0 $$

And also $V(y_i) = \sigma^2$

The above equations show that $\beta$ and $x_i$ are fixed constants while $\epsilon_i$ is a random variable with mean 0 and variance $\sigma^2$. 

\[ y \]

\[ y_1 \]

\[ \epsilon_i = (y_i - \beta x_i) \]

\[ \beta x_i \]

\[ x_1 \]

\[ x_2 \]
As shown in the above figure,

\[ \epsilon_1 = (y_1 - \beta x_1) \] and \[ \epsilon_2 = (y_2 - \beta x_2) \]

so that

\[ \epsilon_1^2 = (y_1 - \beta x_1)^2 \] and \[ \epsilon_2^2 = (y_2 - \beta x_2)^2 \]

thus,

\[ Q = \sum \epsilon_i^2 = \sum (y_i - \beta x_i)^2 \]

To find the minimum of \( Q \) we have

\[ \delta Q = -2 \sum x_i (y_i - \beta x_i) = 0 \]

\[ \delta \beta \]

and so

\[ \sum x_i y_i = \beta \sum x_i^2 \]

Hence

\[ b = \sum x_i y_i, \text{ so that } b \text{ is the least squares estimation of } \beta. \text{ It is the value of } \beta \text{ that } \sum x_i^2 \text{ satisfies this relation.} \]

If we let

\[ x_i = x_i - x \] and \[ y_i = y_i - y \]

then we have

\[ b = \frac{\sum (x_i - x)(y_i - y)}{\sum (x_i - x)^2} = \frac{n \sum xy - (\sum x)(\sum y)}{n \sum x^2 - (\sum x)^2} \]

This line passes through the origin. However, there is not usually any reason why the line that we fit should pass through the origin. More commonly, we want to fit a straight line whose equation is assumed to be

\[ y = \alpha + \beta x \]

so that the observations will have the relationship

\[ y_i = \alpha + \beta x_i + \epsilon_i \]
hence
\[ Q = \sum \hat{e}_i^2 = \sum (y_i - \alpha - \beta x_i)^2 \]
differentiating the above with respect to \( \alpha \) and \( \beta \) one after the other gives
\[ \delta Q = -2 \sum (y_i - \alpha - \beta x_i) = 0 \quad \text{.......................... (1)} \]
\[ \delta \alpha \]
\[ \delta Q = -2 \sum x_i (y_i - \alpha - \beta x_i) = 0 \quad \text{.......................... (2)} \]
\[ \delta \beta \]
and solving the resulting equations simultaneously yields
\[ a = y - bx \quad \text{and} \quad b = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{\sum (x_i - x)(y_i - y)}{\sum (x_i - x)^2} \quad \text{...... (3)} \]
The equations
\[ \sum (y_i - \alpha - \beta x_i) = 0 \]
\[ \sum x_i (y_i - \alpha - \beta x_i) = 0 \]
are called the normal equations.

Dividing both sides by \(-2\) and moving the summation signs inside, we have
\[ \sum y_i - n a - b \sum x_i = 0 \quad \text{................. (4)} \]
\[ \sum x_i y_i - a \sum x_i - b \sum x_i^2 \quad \text{................. (5)} \]
Equation 3(a) can be written as
\[ a = \frac{\sum x_i y_i}{\sum x_i} - \frac{b \sum x_i}{n} \]
These values of the parameters i.e. \( a \) and \( b \) obtained from the data are substituted into the mode called regression equation \( Y = a + bx \).

Prediction can be made within the available data that is within the range of the given data (interpolation) as well as outside the given range (extrapolation).
Types of Regression

There are two types of regression. When $y$ is depending on $x$, that is, when a small change $x$ affects $y$ to change, this is the regression of $y$ on $x$. But when $x$ is depending on $y$, it is called regression of $x$ on $y$. That means that in the first case, $x$ is an independent variable while $y$ is a dependent variable. The two equations are written below:

Regression of $y$ on $x$ is $Y = a + bx$, $a$ is an intercept and $b$ remain the regression coefficient or slope or gradient. For the regression of $x$ on $y$, it is $X = a + by$. We should understand that the two lines will not coincide with the regression equation. The values of $b$ are not going to be the same.

Example: The following data gives the total cost of advertising for a given product and profit made by Chintex Wax textile Company for ten months.

<table>
<thead>
<tr>
<th>Months</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advertising cost in `000</td>
<td>80</td>
<td>90</td>
<td>70</td>
<td>65</td>
<td>82</td>
<td>75</td>
<td>68</td>
<td>76</td>
<td>85</td>
<td>69</td>
</tr>
<tr>
<td>Profit in `000</td>
<td>12</td>
<td>16</td>
<td>9</td>
<td>8</td>
<td>16</td>
<td>10</td>
<td>9</td>
<td>14</td>
<td>13</td>
<td>10</td>
</tr>
</tbody>
</table>

a. Fit a least squares regression line of profit on cost of advertisements. Hence estimate the expected profit when the cost of advertisements are:
   i. `95,000
   ii. `72,000

b. Draw the scatter diagram for the data and comment on the direction of the trend.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$y$</th>
<th>$xy$</th>
<th>$x^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>12</td>
<td>960</td>
<td>6400</td>
</tr>
<tr>
<td>90</td>
<td>16</td>
<td>1440</td>
<td>8100</td>
</tr>
<tr>
<td>70</td>
<td>9</td>
<td>630</td>
<td>4900</td>
</tr>
<tr>
<td>65</td>
<td>8</td>
<td>520</td>
<td>4225</td>
</tr>
<tr>
<td>82</td>
<td>16</td>
<td>1312</td>
<td>6724</td>
</tr>
<tr>
<td>75</td>
<td>10</td>
<td>750</td>
<td>5625</td>
</tr>
<tr>
<td>68</td>
<td>9</td>
<td>612</td>
<td>4624</td>
</tr>
</tbody>
</table>
a.  The regression equation is

\[ Y = a + bx \]

And the normal equations are

\[ \sum y = na + b\sum x \]
\[ \sum xy = a\sum x + b\sum x^2 \]

Substituting in the two equations above to obtain the parameters of the equations.

\[ 117 = 10a + 760b \quad \ldots \ldots \ldots \quad (1) \]
\[ 9083 = 760a + 58360b \quad \ldots \ldots \ldots \quad (2) \]

(1) x 76 gives

\[ 8892 = 760a + 57760b \quad \ldots \ldots \ldots \quad (3) \]

(2) – (3) gives

\[ 191 = 600b \]
\[ b = \frac{191}{600} = 0.3183 \]

Substitute in the value \( b = 0.3183 \) in (1)

\[ 117 = 10a + 760(0.3183) \]
\[ 117 = 10a + 241.908 \]
\[ 10a = 117 - 241.908 \]
\[ 10a = -124.908 \]
\[ a = -124.908 / 10 \]
\[ a = -12.4908 \]

Therefore

\[ Y = -12.4908 + 0.3183x \]
i. Given that $x = 95$, the profit will be

\[ Y = -12.4908 + 0.3183(95) \]
\[ Y = -12.4908 + 30.2385 \]
\[ Y = 17.7477 \]

The profit after spending ₦95,000 on adverts will be ₦17,747.70.

ii. Given that $x = 72$, the profit will be

\[ Y = -12.4908 + 0.3183(72) \]
\[ = -12.4908 + 22.9176 \]
\[ = 10.4268 \]

It means that if ₦72,000 could be spent on advertisement, the profit for the company will be ₦10,426.80.

b. See the scatter diagram below.

The scatter-plot shows that the more the company on adverts, the more the profits.
Correlation

Our analysis so far has been based on the assumption of an existence of a form of relationship between the two variates X and Y. There are times when we are interested in the degree of association between these two variables. This will not always involve calculating the regression line. It will only be a case of defining and calculating a measure of association between the variables.

Correlation is the amount of similarity, in direction and degree, in corresponding pairs of observations of two variables. It is the process of finding how well (or badly) the line fits the observations, such that if all the observations lie exactly on the line of best fit, the correlation is considered to be 1 or unity.

The correlation between two variables reflects the degree to which the variables are related. The most common measure of correlation is the Pearson Product Moment Correlation (called Pearson's correlation for short). When measured in a population the Pearson Product Moment correlation is designated by the Greek letter rho (\( \rho \)). When computed in a sample, it is designated by the letter "r" and is sometimes called "Pearson's r." Pearson's correlation reflects the degree of linear relationship between two variables. It ranges from +1 to -1. A correlation of +1 means that there is a perfect positive linear relationship between variables. The scatter plot shown above depicts such a relationship. It is a positive relationship because high scores on the x-axis are associated with high scores on the y-axis.

The Product Moment Correlation Coefficient

The degree of correlation between X and Y as variables is measured by the product moment correlation coefficient r. The formula for Pearson's correlation takes on many forms. A commonly used formula is shown on the right. The formula looks a bit complicated, but taken step by step as shown in the numerical example, it is really quite simple.

\[
r = \frac{\sum XY - \sum X \sum Y}{\sqrt{\left(\sum X^2 - (\sum X)^2\right)\left(\sum Y^2 - (\sum Y)^2\right)}}
\]
A simpler looking formula can be used in this form:

\[ r = \frac{\sigma_{xy}}{\Sigma x \sigma y} = \frac{\sum(x - m_x)(y - m_y)}{\sum(x - m_x)^2} \]

where \( m_x \) is the mean for variable \( X \) and \( m_y \) is the mean for variable \( Y \).

This can be written as

\[ \rho = \frac{n \sum XY - \sum x \sum y}{\sqrt{[n \sum X^2 - (\sum X)^2][n \sum Y^2 - (\sum Y)^2]}} \]

**Example:** The following are the data on profits of a company in Nigeria and investments made by the company management. The figures for both are in ₦ million.

<table>
<thead>
<tr>
<th>Profits (x)</th>
<th>1.2</th>
<th>2.5</th>
<th>1.5</th>
<th>0.9</th>
<th>0.8</th>
<th>1.5</th>
<th>2.1</th>
<th>1.8</th>
<th>1.2</th>
<th>1.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Investment (y)</td>
<td>1.0</td>
<td>2.4</td>
<td>1.2</td>
<td>0.6</td>
<td>0.3</td>
<td>1.3</td>
<td>1.7</td>
<td>1.6</td>
<td>0.7</td>
<td>1.4</td>
</tr>
</tbody>
</table>

a. Calculate the product moment correlation coefficient.

b. Comment on both the sign and the magnitude of the coefficient.

**Solution**

<table>
<thead>
<tr>
<th>( X )</th>
<th>( y )</th>
<th>( x^2 )</th>
<th>( y^2 )</th>
<th>( xy )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.2</td>
<td>1.0</td>
<td>1.44</td>
<td>1.00</td>
<td>1.2</td>
</tr>
<tr>
<td>2.5</td>
<td>2.4</td>
<td>6.25</td>
<td>5.76</td>
<td>6.0</td>
</tr>
<tr>
<td>1.5</td>
<td>1.2</td>
<td>2.25</td>
<td>1.44</td>
<td>1.8</td>
</tr>
<tr>
<td>0.9</td>
<td>0.6</td>
<td>0.81</td>
<td>0.36</td>
<td>0.54</td>
</tr>
<tr>
<td>0.8</td>
<td>0.3</td>
<td>0.64</td>
<td>0.09</td>
<td>0.24</td>
</tr>
</tbody>
</table>
\[\begin{array}{cccccc}
1.5 & 1.3 & 2.25 & 1.69 & 1.95 \\
2.1 & 1.7 & 4.41 & 2.89 & 3.57 \\
1.8 & 1.6 & 3.24 & 2.56 & 2.88 \\
1.2 & 0.7 & 1.44 & 0.49 & 0.84 \\
1.9 & 1.4 & 3.64 & 1.96 & 2.66 \\
\end{array}\]

\[
\sum x = 15.4 \quad \sum x^2 = 26.34 \quad \sum y = 12.2 \quad \sum y^2 = 18.24 \quad \sum xy = 21.68
\]

\[
\rho = \frac{n\sum XY - \sum x \sum y}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2]}}
\]

\[
r = \frac{10(21.68) - (15.4)(12.2)}{\sqrt{[10(26.34) - (15.4)^2][10(18.24) - (12.2)^2]}} = \frac{216.8 - 187.88}{29.6751} = 0.9745
\]

**Comment:** \( r \) is very close to +1, so there is a very close (linear) relation between \( x \) and \( y \), both increasing together. This means that as the company in question have more profit it increase her investment of such additional fund.

**Spearman’s Rank Correlation Coefficient**

When working with the ordinal scale, it is impossible to correctly measure our experimenter objects since it is impossible to quantify such objects. Generally, ranking of objects occur in such places. In dealing with a correlation problem where the values are in ranks we use rank correlation methods. In statistics, Spearman’s rank correlation coefficient, named after Charles Spearman is a non-parametric measure of correlation- that is, it assesses how well an arbitrary monotonic function could describe the relationship between two variables, without making any assumptions about the frequency distribution of the variables.
Unlike the Pearson product-moment correlation, Spearman’s rank correlation coefficient does not require the assumption that the relationship between the variables is linear, nor does it require the variables to be measured on interval scales; it can be used for variables measured at the ordinal level.

However, Spearman’s rho does assume that subsequent ranks indicate equi-distant positions on the variable measured. For example, using Spearman’s rho for Likert scales often used in psychology, sociology, education and related disciplines assumes that the (psychologically) “felt distances” between scale points are the same for all between of the Likert scale used.

Where equi-distant cannot be justified, correlation between ordinal-level variables can be calculated by using Kendall’s \(\tau\) (tau).

Spearman’s rho is a measure of the linear relationship between two variables. It differs from Pearson’s correlation only in that the computations are done after the numbers are converted to ranks. When converting to ranks, the smallest value on X becomes a rank of 1, etc.

The spearman’s ranked correlation coefficient is defined as follows:

\[
r = 1 - \frac{\sum d^2}{n(n^2 - 1)}
\]

where \(d\) is the difference in each pair of rank

\(n\) is the number of objects being ranked

\(r\) is between -1 and +1

Interpretation of \(r\)

i. When \(r = 1\), it indicates a perfect positive correlation.

ii. When \(r = -1\), it indicates a perfect negative correlation.

iii. When \(-1 < r < -0.5\), it indicates a strong negative correlation.

iv. When \(-0.5 < r < 0\), it indicates a weak negative correlation.

v. When \(0 < r < 0.5\), it indicates a weak positive correlation.

vi. When \(0.5 < r < 1\), it indicates a strong positive correlation.
Example:
The following are the scores of eleven students in two core courses in the faculty

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSS 204</td>
<td>41</td>
<td>37</td>
<td>38</td>
<td>29</td>
<td>49</td>
<td>47</td>
<td>42</td>
<td>34</td>
<td>36</td>
<td>48</td>
<td>29</td>
</tr>
<tr>
<td>ECO 204</td>
<td>36</td>
<td>20</td>
<td>31</td>
<td>24</td>
<td>37</td>
<td>35</td>
<td>42</td>
<td>26</td>
<td>27</td>
<td>29</td>
<td>23</td>
</tr>
</tbody>
</table>

Do the scores show that a student who passed FSS 204 will also pass ECO 204?

<table>
<thead>
<tr>
<th>Student</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rx</td>
<td>5</td>
<td>7</td>
<td>6</td>
<td>10.5</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>9</td>
<td>8</td>
<td>2</td>
<td>10.5</td>
</tr>
<tr>
<td>Ry</td>
<td>3</td>
<td>11</td>
<td>5</td>
<td>9</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>d = Rx - Ry</td>
<td>2</td>
<td>-4</td>
<td>1</td>
<td>1.5</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>-4</td>
<td>0.5</td>
</tr>
<tr>
<td>d²</td>
<td>4</td>
<td>16</td>
<td>1</td>
<td>2.25</td>
<td>1</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[\sum di = 52.5 \quad n = 11\]

\[r = 1 - \frac{6(52.5)}{11(121 - 1)} = 1 - \frac{6(52.5)}{1320} = 1 - 0.2386 = 0.7614\]

There is a strong positive relationship between the scores of these eleven students. It shows that the students have flair for calculation.

Summary
Regression analysis is a forecasting technique used to establish the relationship between quantifiable variables. In regression analysis, data on dependent and independent variables is plotted on a scatter graph and trends are indicated through a line of best fit.

The correlation coefficient provides a very useful summary of the relationship between x and y. But it takes real effort to use a knowledge
of the correlation coefficient and the value of x (or y) to make prediction about the value of y (or x)

Pearson’s sample correlation coefficient, $r$, measures the direction and the strength of the linear association between two numerical paired variables.

Post-Test

1. a. Explain what is meant by a rank correlation coefficient. Outline two circumstances in which a rank correlation coefficient might be used rather than ordinary (i.e. product moment) correlation coefficient.

   b. The table below gives the numbers of ships that birth on our coast per month and the statistics of vehicles brought to the country.

   Post-Test

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>Aug</th>
<th>Sept</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>No of Ships</td>
<td>100</td>
<td>115</td>
<td>117</td>
<td>223</td>
<td>118</td>
<td>230</td>
<td>233</td>
<td>220</td>
<td>250</td>
<td>245</td>
<td>251</td>
<td>95</td>
</tr>
<tr>
<td>No of Vehicles</td>
<td>1500</td>
<td>1002</td>
<td>1345</td>
<td>1333</td>
<td>1300</td>
<td>1330</td>
<td>1250</td>
<td>1233</td>
<td>1266</td>
<td>1333</td>
<td>1200</td>
<td>1170</td>
</tr>
</tbody>
</table>

   a. Draw a scatter diagram of the data and identify the outlier.

   b. Calculate Spearman’s rank correlation coefficient for the data.

   c. Recalculate the correlation coefficient excluding the outlier. Comment briefly on the two coefficients.

2. LASMA Ferries runs ferries to mile 2 from Marina. The data below give the prices (in N) for a return ticket for a driver and for a car on each ten routes.

<table>
<thead>
<tr>
<th>Route</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
</tr>
</thead>
<tbody>
<tr>
<td>Driver cost (x)</td>
<td>20</td>
<td>23</td>
<td>27</td>
<td>33</td>
<td>28</td>
<td>42</td>
<td>38</td>
<td>23</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>Car cost (y)</td>
<td>92</td>
<td>107</td>
<td>124</td>
<td>165</td>
<td>105</td>
<td>163</td>
<td>143</td>
<td>85</td>
<td>100</td>
<td>83</td>
</tr>
</tbody>
</table>
You are given that

\[ \sum x^2 = 8113, \quad \sum y^2 = 144671, \sum xy = 34046. \]

1. Plot a scatter diagram of the data, marking the letters near your points.
2. Find the correlation coefficient between x and y and comment on its value.
3. Find the regression line that predicts car cost for a given driver cost.
4. Plot the line on your scatter diagram. Which route gives the cheapest actual car cost compared to predicted cost and which is the most expensive?

3. a. Explain briefly what the product–moment correlation coefficient, r, of two variables measures.

   b. Draw a sketch of a possible scatter plot of two variables x and y when
      i. \( r = +1 \)
      ii. \( r = -1 \)
      iii. \( r = 0 \)

   c. Give one similarity and one difference between the product–moment correlation coefficient and Spearman’s rank correlation coefficient, r.

   d. Ten Contractors out of those who sought for loan to execute some projects were picked at random and sent to two finance outfits to recommend the amount of money to be given as loan to each of the applicants. Based on their estimates, and other criteria, the two finance outfits A and B recommended as follows:

<table>
<thead>
<tr>
<th>Applicant</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>211</td>
<td>282.50</td>
<td>287.50</td>
<td>262.80</td>
<td>274.10</td>
<td>237.60</td>
<td>282.50</td>
<td>258.40</td>
<td>227</td>
<td>281.70</td>
</tr>
<tr>
<td>B</td>
<td>268.5</td>
<td>294.10</td>
<td>288.40</td>
<td>278.90</td>
<td>281.70</td>
<td>270.70</td>
<td>288.60</td>
<td>274</td>
<td>245</td>
<td>281.70</td>
</tr>
</tbody>
</table>

Calculate spearman’s rank correlation coefficient for the data. Would you say that the two banks agree in their assessment of the applicants’ needs? Why?
e. State one major demerit of rank correlation and discuss briefly how this is confirmed in the problem above.

4. a. What is a scatter diagram?
   b. The results of sales before and after adverts were observed by the Marketing Manager of ID Nobles Nigeria Limited. The following were his observations in ten months:

<table>
<thead>
<tr>
<th>Sales</th>
<th>Quantity in tones</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before adverts</td>
<td>45 50 85 20 66 78 100 90 116 110</td>
</tr>
<tr>
<td>After adverts</td>
<td>32 80 80 60 80 85 98 92 150 125</td>
</tr>
</tbody>
</table>

   i. Construct the scatter diagram and comment on its feature.
   ii. Use the method of least-squares to determine an equation for the best straight line for these data. Interpret the meaning of the slope in this equation.
   iii. How many quantities in tons would you predict that 80 tonnes of the item will be sold after adverts?

References


LECTURE NINE

Theory of Probability

Introduction
Apart from the fact that the theory of probability was developed with the study of games and chance, uncertainty prevailed in every sphere of life. Take for instance; one broker often predicts that the stock price will surely improves next month. It is quite likely that there will be a good yield of corn this year and so on. This indicates that the word probability connotes that there is an uncertainty about the happening of events. To put probability on a better footing we define it, but before doing so, let us state the objectives of the lecture.

Objectives
At the end of this lecture, you should be able to:
1. define probability;
2. state its concepts;
3. state the properties of probability;
4. state the laws of probability and apply them to real life situations; and
5. define and apply conditional probability.

Pre-Test
1. The incidence of occupational disease in an industry is such that the workmen have 20% chance of suffering from it. What is the probability of 4 or more workmen out of 6 contacting the disease?
2. A community of 12 landlords consists of 4 high income earners, 5 low income earners and 3 middle income earners. A committee of three landlords is to be chosen to represent the community. What is the chance that:
   i. the three landlords belong to different income group;
   ii. two belong to one class and the third to different class;
   iii. the three belong to the same class.
3. Explain the following statistical terms:
   i. independent events
   ii. sample space
   iii. mutually exclusive events
   iv. sample point
4. There are three children in a family, what is the probability that they include:
   i. exactly two girls
   ii. not more than one girl?
5. Two students are chosen at random from a class of 22 boys and 18 girls. Find the probability that two students selected are of the same sex.
6. a. What is Permutation?
   b. A committee of 7 representative of a class consists of a class representative and his deputy. On a visit to the Head of department, there are four seats available. How many ways can the committee be seated if:
      i. there is no restriction
      ii. the class rep. and his deputy must sit together
      iii. one of the students has committed a crime and cannot sit down even if there were enough seats
7. What is probability? Explain the following concepts of probability:
   i. Classical probability;
   ii. Axiomatic probability
   iii. Relative frequency
CONTENT

Meaning of Probability

Probability is a measure of degree of believes in the occurrence of an event.

Other Definitions

The following are the definitions of probability:

i. Probability is a mathematical term which shows the degree of belief we have about something.

ii. It is the science of decision making with calculated risks in the face of uncertainty.

iii. Probability is a measure of the likelihood of a required event happening.

The concept of probability is necessary when dealing with physical, biological, or social mechanisms that generate observations that cannot be predicted with certainty.

The following are the basic concepts in probability:

i. **Random experiment** – Random experiment is an act whose outcome cannot be predetermined e.g. tossing a coin and selection of ball from a box.

ii. **An Experiment** means performing an act e.g. dipping hand into a bowl containing balls/paper, tossing a coin and throwing a dice.

iii. **A trial** is just one act performed. It is a procedure or an experiment to collect any statistical data such as rolling a dice or flipping a coin.

iv. **An outcome** is one of the possible results that can happen in a trial of an experiment e.g., tossing a coin we have head (H) or tail (T) as the outcome.

v. **An event** consists of one or more possible outcomes of an experiment. E.g. An even number in a throw of a die. A = {2, 4, 6}. It is also a collection of outcomes (sample point) which have certain quality in common.

vi. **A sample space** is the collection of all the possible outcomes of an experiment denoted by

\[ S = \{ e_1, e_2, e_3, \ldots , e_n \} \]
e.g. tossing a coin twice, the sample space is

$$\Omega = S = \{HH, HT, TH, TT\}$$

It is the totality of all the outcomes or results of a random experiment.

**Types of sample spaces**

- A *countable* number of events (can be finite or infinite). For example, the sample space for flipping a coin twice is

  $$S = \{(H,H), (H,T), (T,H), (T,T)\}.$$  

- The sample space includes a range of values. For example, suppose a final exam is given from 2:00-4:00 p.m. The sample space for the time a student hands in the exam is

  $$S = \{t : 2:00 \leq t \leq 4:00\}$$

**NOTE:** By countable, we mean either a *finite* or *infinite* number of elements that can be counted. For example, \{1, 2, 3, 4, 5\} is a finite countable set and \{1, 2, 3, \ldots\} is an infinite countable set. Any infinite sequence that we can associate with the numbers \{1, 2, 3, \ldots\} is also (infinitely) countable, e.g., \{1, 3, 5, \ldots\} or \{0, 2, 4, \ldots\}. Note that the sample space in the second example is *uncountable*.

vii. **Equally likely events**: All possible results of a random experiment are called equally likely outcomes and we have no reason to expect any one rather than the other.

viii. **Mutually Exclusive events**: Two events are said to be mutually exclusive if they cannot happen together or the probability of occurring together is zero. These are disjoint, they do not have any element in common and consequently cannot occur simultaneously, i.e., \(P(\text{AnB}) = \emptyset\). If A and B are mutually exclusive events, then \(P(A \text{ or } B) = P(A) + P(B)\).

ix. **Exhaustive Events**: Events are exhaustive when they include all the possibilities associated with the same trial. In throwing a coin, the turnings up of head and of a tail are exhaustive events assuming of course that the coin cannot rest on its edge.

167
Independent Events: Two events are said to be independent if the occurrence of any event does not affect the occurrence of the other event. For example, in tossing of a coin, the events corresponding to the two successive tosses of it are independent. Another example, Let A = event of an odd number in a die and B = event of a head in a coin. These events are independent of each other.

Mrs Jackson is expecting a baby. The probability that it will be a boy is $\frac{1}{2}$ and the probability that the baby will have blue eyes is also $\frac{1}{4}$ . What is the probability that Mrs Jackson will give birth to a blue eyed boy?

Solution:

\[ P(\text{a blue eyed boy}) = \frac{1}{4} \cdot \frac{1}{2} = \frac{1}{8} . \]

Dependent Events: If the occurrence or non-occurrence of any event affects the happening of the other, then the events are said to be dependent events.

Law of the complement: The probability that an event A will not occur is $A'$, which is $1 - P(A) = P(\text{not } A)$.

Example: In an experiment, the probability of an event occurring is 65%. What is the probability of its not occurring?

Solution: Let $A =$ probability of an event occurring. The $P(\text{not occurring}) =$

\[ 1 - p(A) = 1 - \frac{65}{100} = 35\% = 0.35. \]

Other Definitions of Probability

We shall now consider three more definitions of probability.

i. Relative Frequency Approach: defines probability as related to frequency or limits of relative frequency is an indefinitely long sequence of trials i.e.

\[ \text{Probability} = \lim \frac{\text{No of success}}{\text{Total no of trials}} \]
ii. **Classic Definition**

Suppose that an event E can happen in $h$ ways out of a total of $n$ possible equally likely ways. Then the probability of occurrence of the event (called its success) is denoted by

$$p = \Pr(E) = \frac{h}{n}$$

The probability of non-occurrence of the event (called its failure) is denoted by

$$q = \Pr(\text{not } E) = \frac{n - h}{n} = 1 - p = 1 - \Pr(E)$$

Thus $p + q = 1$

iii. **Axiomatic Approach** : Probability of an event $A$ is defined as $P(A)$ such that it satisfied the following axioms:

i. $0 \leq P(A) \leq 1$; that is, the probability of an event $A$ must be greater or equal to zero but less than or equal to unity.

ii. $\sum P(A_i) = P(S) = 1$, $\forall \ i \neq j$

iii. If $A_1, A_2, A_3, \ldots, A_n$, is a finite collection of pairwise mutually exclusive events of $S$ when $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$.

**Addition Rule:**

If events $A$ and $B$ are any two events, the probability of event $A$ or $B$ occurring is defined as:

$$P(A \cup B) = P(A) + P(B) - P(\text{AnB})$$

The probability of happening of any one of the two mutually exclusive events is equal to the sum of their individual probabilities, symbolically,

$$P(A \cup B) = P(A) + P(B)$$

**Multiplication Rule:**

The probability that two independent events $A$ and $B$ occur together is the product of their respective probabilities, that is, $P(A \text{ and } B) = P(A) \cdot P(B)$
Examples

1. Four Directors (A, B, C, D) of a company are being considered as candidates for a two member delegation to represent the company at an international meeting. What is the probability that the director A or director B is selected?

Solution
The sample space \( S = \{AB, AC, AD, BC, BD, CD\} \)
To determine the probability that director A or director D will be selected, let \( M \) be the event that director A is selected and \( N \) be the event that director D is selected, then

\[
M = \{AB, AC, AD\} \\
N = \{AD, BD, CD\} \\
M \cap N = \{AD\}.
\]

Thus, by the rule of addition, the probability of the event \( M \) or \( N \) (that director A or director D is selected) is

\[
P(M \text{ or } N) = P(M) + P(N) - P(M \text{ and } N) \\
= \frac{3}{6} + \frac{3}{6} - \frac{1}{6} \\
= \frac{5}{6}.
\]

2. In a survey to know the HIV/AIDS status of the inhabitants of a community in Benue state, two men are drawn, one at a time, at random, from suspected AIDS careers. What is the probability that:
   i. Two of them are HIV/AIDS positive?
   ii. Both of them are not affected?
   iii. One of them is a career and the other is not affected?

Solution
Let \( D \) represents HIV/AIDS career and \( N \) represents HIV/AIDS negative. Hence the sample space will be

\[
S = \{NN, ND, DN, DD\}
\]
i. Two of them are HIV/AIDS positive. Only one sample point is favourable to the event. Therefore, \( P(DD) = \frac{1}{4} \).

ii. Both of them are not affected. There is one sample point favourable to the event. \( P(NN) = \frac{1}{4} \).

iii. One of them is a career and other is not. There are two sample points favourable to the event. \( P(ND, DN) = \frac{2}{4} = \frac{1}{2} \).

3. An investment consultant predicts that the odds against the price of a certain stock will go up during the next week are 2:1, and the odds in favour of the price remaining the same are 1:3. What is the probability that the price of the stock will go down during the next week? Also, calculate the probability that stock price will go down.

**Solution:**
Let \( A \) denotes the event “stock price will go up”, and \( B \) be the event “stock price will remain same”. Then

\[
P(A) = \frac{1}{3} \quad \text{and} \quad P(B) = \frac{1}{4}
\]

P(stock price will either go up or remain same)

\[
P(A \cup B) = P(A) + P(B) = \frac{1}{3} + \frac{1}{4} = \frac{7}{12}.
\]

Now \( P(\text{stock price will go down}) \)

\[
P(A' \cap B') = 1 - P(A \cup B)
\]

\[
= 1 - \frac{7}{12} = \frac{5}{12}
\]

4. Two dice are rolled. Find the probability that the score on the second die is greater than the score on the first.

**Solution**
**Experiment:** Two dice are rolled. The Large sample space is:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1,1</td>
<td>1,2</td>
<td>1,3</td>
<td>1,4</td>
<td>1,5</td>
<td>1,6</td>
</tr>
<tr>
<td>2</td>
<td>2,1</td>
<td>2,2</td>
<td>2,3</td>
<td>2,4</td>
<td>2,5</td>
<td>2,6</td>
</tr>
<tr>
<td>3</td>
<td>3,1</td>
<td>3,2</td>
<td>3,3</td>
<td>3,4</td>
<td>3,5</td>
<td>3,6</td>
</tr>
<tr>
<td>4</td>
<td>4,1</td>
<td>4,2</td>
<td>4,3</td>
<td>4,4</td>
<td>4,5</td>
<td>4,6</td>
</tr>
<tr>
<td>5</td>
<td>5,1</td>
<td>5,2</td>
<td>5,3</td>
<td>5,4</td>
<td>5,5</td>
<td>5,6</td>
</tr>
<tr>
<td>6</td>
<td>6,1</td>
<td>6,2</td>
<td>6,3</td>
<td>6,4</td>
<td>6,5</td>
<td>6,6</td>
</tr>
</tbody>
</table>
**Event A**: The score on the second die is greater than the score on the first die.

i.e. \( A = \{ (1,2), (1,3), (1,4), (1,5), (1,6), (2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), (5,6) \}. \)

\[ \therefore (A) = 15, \quad n(S) = 6 \times 6 = 36 \]

Therefore, \( p(A) = \frac{n(A)}{n(S)} = \frac{15}{36} = \frac{5}{12} = 0.42 \)

5. The production capacities for the three refineries Port Harcourt, Warri and Kaduna are 0.2, 0.3 and 0.5 of the total fuel being consumed in Nigeria. It was discovered that there are some that are adulterated. The percentage of adulterated produced by Port Harcourt, Warri and Kaduna is 1, 2 and 4 percent respectively. Some kegs of fuel were picked randomly, what is the probability that it is adulterated?

**Solution**

The adulterated fuel produced by Port Harcourt = \( \frac{0.2 \times 1}{100} = 0.002 \)

The adulterated fuel produced by Warri = \( \frac{0.3 \times 2}{100} = 0.006 \)

The adulterated fuel produced by Kaduna = \( \frac{0.5 \times 4}{100} = 0.02 \)

The total percentage of adulterated fuel by all these three refineries = \( \frac{0.2 + 0.6 + 2}{100} = 0.028 \)

6. Some Social Scientists were asked to comment on the economic situation of this country. Three of them (labeled A, B, C) were randomly picked. They stated that the chances of having improved
economy are \( \frac{1}{2}, \frac{3}{4}, \) and \( \frac{1}{4} \) respectively. What is the probability that the economy situation in the country will be improved? Also calculate the probability that the economy situation will not improve.

**Solution**
The probability that the problems can be solved as expressed by these three social scientists is \( \frac{1}{2} \times \frac{3}{4} \times \frac{1}{4} = \frac{3}{32} \).

The probability that it cannot be solved is \( 1 - \frac{3}{32} = \frac{29}{32} \).

7. Ayo and Tayo appear in an interview for two vacancies in the same post. The chance of Ayo’s selection is \( \frac{1}{7} \) and that of Tayo’s selection is \( \frac{1}{5} \). What is the probability that
   i. Both of them will be selected?
   ii. Only one of them will be selected?
   iii. One of them will be selected?

**Solution**
Let A represents Ayo and T represents Tayo
   i. \( P(A \text{ selection}) = \frac{1}{7} \) and \( P(B \text{ selection}) = \frac{1}{5}, \)

   \[ P(\text{both selected}) = \frac{1}{7} \times \frac{1}{5} = \frac{1}{35} \]

   ii. \( P(\text{only one is selected}) = P(\text{only A is selected and T not selected}) \)

   \[ = \{\frac{1}{7} \times \frac{4}{5}\} + \{\frac{1}{5} \times \frac{6}{7}\} = \frac{10}{35} = \frac{2}{7} \]

   iii. \( P(\text{none of them will be selected}) = \frac{6}{7} \times \frac{4}{5} = \frac{24}{35} \)

**Example 8:**
The probability that two independent companies A and B will survive in the next ten years are \( \frac{3}{4} \) and \( \frac{4}{5} \) respectively. Calculate the probabilities that in the next ten years,
   a. Both companies will survive
   b. None of them will survive
c. One of them will survive
d. At least one of them will survive

Solution

\[ P(A) = \frac{3}{4} \quad P(A^\prime) = 1 - \frac{3}{4} = \frac{1}{4} \]
\[ P(B) = \frac{4}{5} \quad P(B^\prime) = 1 - \frac{4}{5} = \frac{1}{5} \]

a. \[ P(AnB) = P(A) \cdot P(B) = \frac{3}{4} \cdot \frac{4}{5} = \frac{3}{5} \]
b. \[ P(AnB^\prime) = P(A) \cdot P(B^\prime) = \frac{1}{4} \times \frac{1}{5} = \frac{1}{20} \]
c. \[ P[(AnB) \cup (AnB^\prime)] = P(AnB) + P(AnB^\prime) \]
\[ = P(A) \cdot P(B) + P(A) \cdot P(B^\prime) \]
\[ = \frac{3}{4} \times \frac{4}{5} + \frac{1}{4} \times \frac{4}{5} \]
\[ = \frac{3}{20} + \frac{4}{5} \]
\[ = 7/20 \]
d. \[ P(x \geq 1) = 1 - P(x = 0) \]
\[ = 1 - P(AnB) \]
\[ = 1 - \frac{1}{20} = \frac{19}{20}. \]

Conditional Probability

Let A and B be two events, the probability that an event A occurs given that B has occurred is the conditional probability of A given B. It is defined as

\[ P(A/B) = \frac{P(AnB)}{P(B)} \]

provided \( P(B) \neq 0 \) \( \cdots \cdots (1) \)

From (1),

\[ P(AnB) = P(B) \cdot P(A/B) \]

The probability of B given that event A has occurred is defined as

\[ P(B/A) = \frac{P(AnB)}{P(A)} \]

provided \( P(A) \neq 0 \) \( \cdots \cdots (3) \)
From (3),
\[ P(AnB) = P(A) \cdot P(B/A) \] \hspace{2cm} (4)

Equations (2) and (4) give the multiplication law for conditional probability. That is, \( P(AnB) = P(A) \cdot P(B) \). This shows that the two events are independent.

**Example 7**
The breakdown of the ministers in the country shows that 35\% are female. Out of the female ministers, 20\% of them are PDP members while 40\% of the males are PDP members.

Find the probability that a minister chosen at random from the cabinet is

i. A female PDP member.

ii. A male but not a PDP member.

**Solution**
Let \( F \) represents female minister
Let \( M \) represents male minister
Let \( D \) represents PDP member

Then
\[ P(F) = 35\% \text{ or } 0.35, \quad P(D/F) = 20\% \text{ or } 0.20 \]
\[ P(M) = 1 - 35\% = 65\% = 0.65 \]
\[ P(D/M) = 40\% \text{ or } 0.40 \quad P(D'/M) = 1 - 40\% = 60\% \text{ or } 0.60 \]

\begin{align*}
P(D/M) &= 40\% \text{ or } 0.40 \quad P(D'/M) = 1 - 40\% = 60\% \text{ or } 0.60 \\
\end{align*}

i. A female PDP member = \( P(FnD) = P(F) \cdot P(D/F) \)
\[ = 0.35 \times 0.2 \]
\[ = 0.07 \]

ii. A male but not a PDP member = \( P(MnD') = P(M) \cdot P(D/M) \)
\[ = 0.65 \times 0.6 = 0.39 \]
Example 8
A group of executives are classified according to the status of the body weight and incidence of hypertension. The proportions in the various categories are as shown in the table below:

<table>
<thead>
<tr>
<th>Category</th>
<th>Overweight</th>
<th>Normal weight</th>
<th>Underweight</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypertensive</td>
<td>0.10</td>
<td>0.08</td>
<td>0.02</td>
<td>0.20</td>
</tr>
<tr>
<td>Not-hypertensive</td>
<td>0.15</td>
<td>0.45</td>
<td>0.20</td>
<td>0.80</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>0.25</strong></td>
<td><strong>0.53</strong></td>
<td><strong>0.22</strong></td>
<td><strong>1.00</strong></td>
</tr>
</tbody>
</table>

a. What is the probability that a person selected at random from the group will have hypertension?
b. What is the probability that a person selected from the group who is found to be overweight will also have hypertension?

Solution
Let A = event that a person is hypertensive
B = event that a person is overweight

A = T1A + T2A + T3A

\[ P(A) = P(\text{AT1 or AT2 or AT3}) \quad \text{[inclusive ‘or’]} \]

\[ = P(\text{AT1}) + P(\text{AT2}) + P(\text{AT3}) \]

\[ = 0.10 + 0.08 + 0.02 \]

\[ = 0.20 \]

\[ P(A/B) = P(\text{AnB}) = \frac{0.10}{0.40} = 0.40 \]

P(B) \quad 0.25

Tree Diagram
This is a tree like diagram used in solving probability questions. It is a set of connected lines with each line looking like a branch of a tree. It is a device to represent the trials or stages of an experiment in a diagram.
**Example:** A committee is made up of two males and three females. Two members are selected successively at random without replacement. Draw a tree diagram to bring out these details and list out the sample points.

![Tree Diagram]

The following are the outcomes and their corresponding probabilities:
- **MM** – $\frac{2}{5} \times \frac{1}{4} = \frac{1}{10}$.
- **MF** – $\frac{2}{5} \times \frac{3}{4} = \frac{3}{10}$
- **FM** – $\frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$
- **FF** – $\frac{3}{5} \times \frac{2}{4} = \frac{3}{10}$.

**Counting Techniques; Permutations and Combinations**

The basic rule for counting is the multiplication rule, which we will refer to as the basic counting principle. It can be stated thus:

If a certain experiment can be performed in $r$ ways, and corresponding to each of these ways another experiment can be performed in $k$ ways, then the combined experiment can be performed in $rk$ ways.

It is assumed that the trial of such experiment has a finite sample space and the experiment is repeated $n$ times.

For example, if a coin is thrown three times, we have $2^3 = 8$ outcomes.

$S = \{HHH, HHT, HTH, THH, THT, TTH, HTT, TTT\}$.

**Permutation**

This is the ordered arrangement of $n$ distinct objects taking $r$ at a time. This is written as
\[ \text{nPr = } \frac{n!}{(n-r)!} = n(n-1)(n-2)(n-3) \ldots (n-r+1) \]

In particular, the number of permutations of \( n \) objects taken \( n \) at a time is
\[ \text{nPn} = n(n-1)(n-2) \ldots 1 = n! \]
that is taken all together.

**For example**

Five people arrive at the checkout counter at the same time. In how many different ways can these people line up?

**Solution**
The number of ways of arranging five people in a line is five factorial, that is,
\[ 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120 \]
The number of permutations of \( n \) objects consisting of groups of which \( n_1 \) are alike, \( n_2 \) are alike, \ldots is
\[ \frac{n!}{n_1!n_2!} \]
where \( n = n_1 + n_2 \)

**Example:**
The number of permutations of letters in the word STATISTICS is

Since there are 3s, 3t, 2i, 1a and 1c, hence, the number of ordered arrangement is
\[ \frac{10!}{3!3!2!1!1!} = 50,400 \text{ ways.} \]

**Example:** In how many ways can 10 people be seated on a bench if only 4 seats are available?
Solution
The number of arrangements of 10 people taken 4 at a time is

\[
\frac{10!}{(10-4)!} = 10 \times 9 \times 8 \times 7 = 5040 \text{ ways}
\]

Combinations
A combination of \( n \) different objects taken \( r \) at a time is a selection of \( r \) out of \( n \) objects with no attention given to the order of arrangement. This is denoted

\[
\binom{n}{r} = \frac{n(n-1)\ldots(n-r+1)}{r!} = \frac{n!}{r!(n-r)!}
\]

Example: In how many ways can 10 objects be split into two groups containing 4 and 6 objects respectively?

Solution
This is the same as the number of arrangements of ten objects of which four objects are alike and 6 other objects are alike.

\[
\frac{10!}{4!6!} = \frac{10 \times 9 \times 8 \times 7}{4 \times 3 \times 2 \times 1} = 210
\]

Summary
In this lecture, you have learnt probability and its applications to human behaviour. By now, the fear of what probability is, is gone, you can now apply the concepts so learnt to your course of study.

Post-Test
1. The incidence of occupational disease in an industry is such that the workmen have 20% chance of suffering from it. What is the probability of 4 or more workmen out of 6 contacting the disease?
2. A community of 12 landlords consists of 4 high income earners, 5 low income earners and 3 middle income earners. A committee of three landlords is to be chosen to represent the community. What is the chance that:
   i. the three landlords belong to different income group;
   ii. two belong to one class and the third to different class;
   iii. the three belong to the same class.
3. Explain the following statistical terms:
   i. independent events
   ii. sample space
   iii. mutually exclusive events
   iv. sample point
4. There are three children in a family, what is the probability that they include:
   i. exactly two girls
   ii. not more than one girl?
5. Two students are chosen at random from a class of 22 boys and 18 girls. Find the probability that two students selected are of the same sex.
6. a. What is Permutation?
   b. A committee of 7 representative of a class consists of a class representative and his deputy. On a visit to the Head of department, there are four seats available. How many ways can the committee be seated if:
      i. there is no restriction
      ii. the class rep. and his deputy must sit together
      iii. one of the students has committed a crime and cannot sit down even if there were enough seats
7. What is probability? Explain the following concepts of probability:
   i. Classical probability;
   ii. Axiomatic probability
   iii. Relative frequency
References


LECTURE TEN

Probability Distributions

Introduction
A listing of the values taken by a random variable and their associated probabilities is a probability distribution. It is called the theoretical counterpart of a frequency distribution. There are discrete and continuous probability distributions. But before we discuss these fully, let us highlights what we shall come across in this lecture.

Objectives
At the end of this lecture, students are expected to have mastered the following:

1. the meaning of random variables and types;
2. properties of probability function;
3. binomial distribution;
4. poisson distribution; and
5. normal distribution.

Pre-Test
1. The mean annual income of a worker in a certain professional category is ₦232,000 with a standard deviation of ₦14,000. Assuming that the distribution of income is approximately normal, what is the probability that two workers picked at random will each earn less than ₦225,000?
2. A new vaccine was tested on 100 persons to determine its effectiveness. If the claim of the drug company is that a random
person who is given the vaccine will develop immunity with probability 0.8, find the probability that:

a. Less than seventy-four people will develop immunity
b. Between seventy-four and eighty-five people, inclusive, will develop immunity.

Hint: \( \mu = np \) and \( \sigma = \sqrt{npq} \)

3. If the probability that a child is a son is 0.4, find the probability that in a family of four children there are:
   i. Two sons
   ii. At least two sons
   iii. All girls
   iv. Three girls

4. A machine produces defective items at the rate of 1 in 10. If 100 items produced by this machine are inspected, what is the expected number of defective items? Also find the variance of the number of defective items.

5. Suppose a company received a consignment of 200 boxes of fluorescent tubes, eight tubes to a box. If the probability that a tube is defective is 0.1, how many of the boxes would you expect to have three defective tubes?

6. Give the expression for the probability distribution of a binomial variable \( x \) with parameters \( n \) and \( p \) where \( n \) is the total number of independent trials and \( p \) the probability of a “success” in a trial.

   a. What does \( x \) represent?
   b. State the mean and standard deviation of the variable.
   c. If the variable \( x \) were to be approximated to a Poisson variate with parameter \( \lambda \), state the condition under which this can be done. What would then be the approximate probability distribution of \( x \)? Give the mean and standard deviation of the distribution.

7. Give the following characteristics of a binomial distribution \( f(x) \) with parameter \( p \) and number of independent trials \( n \).

   i. Probability function \( f(x) \)
   ii. Mean
   iii. Variance and its standard deviation.
8. The probability that an entering university student will graduate is 0.4. Determine the probability that of 5 students
   i. None will graduate
   ii. One will graduate
   iii. At least one will graduate
9. The probability that goods produced by a factory are defective is 0.002. If 2000 of such goods are produced, what is the probability that:
   i. None will be defective
   ii. Exactly one will be defective
   iii. Less than two will be defective.
10. a. State the mathematical expressions for the probability of the:
   i. normal distribution
   ii. binomial distribution
   iii. Poisson distribution
   b. Write the binomial expansion of
      i. $(p + q)^4$
      ii. Hence expand $(5/6 + 1/6)^5$

CONTENT

Concept of Random Variable

A random variable assigns numerical values to the outcomes of a chance experiment. Mathematically, a random variable is a function defined on the outcomes of the sample space. In the population, the values of the random variable may be distributed according to some definite probability law which can be expressed mathematically on the basis of theoretical considerations and the corresponding probability distribution is known as theoretical probability distribution. For these distributions, a random experiment is theoretically assumed to serve as a model and the probabilities are given by a function of the random variable called probability function.

The probability distribution herein described as theoretical distribution are useful in that they are scientific way of drawing inferences about the population characteristics. They also serve as
benchmarks against which to compare the actual frequency distributions and to find out whether the difference is due to fluctuations of sampling or some other causes. They can be taken as substitutes for actual distributions when latter cannot be obtained at all. This theoretical distribution helps in taking decisions on the face of uncertainty. They also help in forecasting. Many businesses related problems can be solved on the basis of theoretical distributions.

**Properties of a Probability Function**

i. The probability that a random variable assumes a value $x_i$ is always between 0 and 1. That is, $0 \leq p(x_i) \leq 1$.

ii. The sum of all the probabilities $p(x_i)$ is equal to 1. That is, $\sum_{i=1}^{n} p(x_i) = 1$

**Discrete Random Variables**

The probability distribution of a discrete random variable can be described by listing all the values that the random variable can take, together with the corresponding probabilities. Such a listing is called the probability function of a random variable. This can be in this form

<table>
<thead>
<tr>
<th>Value $x$</th>
<th>Probability $p(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x_1$</td>
<td>$p(x_1)$</td>
</tr>
<tr>
<td>$x_2$</td>
<td>$p(x_2)$</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>$x_n$</td>
<td>$p(x_n)$</td>
</tr>
<tr>
<td><strong>Sum</strong></td>
<td>$\sum p(x_i) = 1$</td>
</tr>
</tbody>
</table>

**Example:**

The number of telephone calls received in the office of Budget and Planning between 12.00 noon and 1.00 p.m has the probability function given as below:
a. Verify that it is in fact a probability function.
b. Find the probability that there will be three or more calls.
c. Find the probability that there will be an even number of calls.

### Solution

a. Since all the probabilities in column 2 are between 0 and 1, and since their sum is 1, we have a genuine probability function.

b. Writing $X$ for the random variable “the number of telephone calls,” we want $P(X = 3)$.

\[
P(X = 3 \text{ or } 4 \text{ or } 5 \text{ or } 6) = p(3) \text{ or } p(4) \text{ or } p(5) \text{ or } p(6)
\]

\[
= 0.20 + 0.10 + 0.15 + 0.05
\]

\[
= 0.5
\]

c. Here we want $P(X \text{ is even})$

\[
P(X \text{ is even}) = p(0) + p(2) + p(4) + p(6)
\]

\[
= 0.05 + 0.25 + 0.10 + 0.05
\]

\[
= 0.45
\]

### The Mean of a Discrete Random Variable

The mean of the probability distribution of a discrete random variable, or simply, the mean of a discrete random variable, is a number obtained by multiplying each possible value of the random variable by the corresponding probability and then adding these terms. Writing $\mu_X$ for the
mean of a random variable $X$, or simply $\mu$ if the context is clear, in the case of a discrete random variable we have the following.

If $x_1, x_2, \ldots, x_n$ are the values assumed by a random variable with respective probabilities $p(x_1), p(x_2), \ldots, p(x_n)$, then its mean $\mu$ is given by

$$
\mu = x_1 \cdot p(x_1) + x_2 \cdot p(x_2) + \ldots + x_n \cdot p(x_n) = \sum_{i=1}^{n} x_i \cdot p(x_i).
$$

the variance of a discrete random variable $X$ is given as

$$
\text{Var}(X) = \sum (x_i - \mu)^2 \cdot p(x_i).
$$

The following are examples of discrete random variable distributions: Binomial and Poisson. Normal distribution is an example of continuous distribution. We shall treat the distributions one by one.

**The Binomial Distribution**

Binomial distribution or Bernoulli distribution is the most fundamental and important discrete probability distribution in statistics. Consider a basic experiment in which the possible outcomes can be classified into one of two categories, a success or a failure. If $p$ is the probability that an event will happen in any single trial (called the probability of a success) and $q = 1 - p$ is the probability that it will fail to happen in a single trial (called the probability of a failure), then the probability that the event will happen $X$ times in $n$ trials (that is, $X$ successes and $n-x$ failures will occur) is given by

$$
p(X) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \quad x = 0, 1, 2, \ldots
$$

If $n$ and $p$ are constants, the above function $p^x$ represents a discrete probability distribution given by:

<table>
<thead>
<tr>
<th>$x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>...</th>
<th>$n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p(x)$</td>
<td>$q^n$</td>
<td>$nC_1 p q^{n-1}$</td>
<td>$nC_2 p^2 q^{n-2}$</td>
<td>$nC_3 p^3 q^{n-3}$</td>
<td>...</td>
<td>$p^n$</td>
</tr>
</tbody>
</table>

187
This is called Binomial distribution because for \( x = 0, 1, 2, \ldots, n \) it gives the successive terms of the binomial expansion:

\[
(q + p)^n = q^n + nC_1 p q^{n-1} + nC_2 p^2 q^{n-2} + \ldots + p^n.
\]

Clearly written \( \sum_{x=0}^{n} p(x) = (q + p)^n = 1 \)

This formula for the binomial distribution assumes that the events:
1. are dichotomous (fall into only two categories)
2. are mutually exclusive
3. are independent and
4. are randomly selected

The two independent constants \( n \) and \( p \) are called parameters of the binomial distribution and the distribution is completely determined, that is, all the probabilities can be obtained if \( n \) and \( p \) are known.

The binomial probability distribution is a discrete distribution. In performing the trials, a tacit assumption is that the outcome of one trial does not influence the outcome of any other trial. Thus, we say that the trials are independent. This assumption, together with other characterizing a binomial distribution can be stated as follows:

i. The binomial experiment consists of a fixed number of trial ‘\( n \)’.
ii. There are only two mutually exclusive outcomes, viz., success or failure for each trial.
iii. The trials are independent.
iv. The probability of success is known and remains the same from trial to trial.

Furthermore, if \( p(\text{success}) = p \) and \( p(\text{failure}) = q \), then \( p + q = 1 \).

**Example 1**: Find the probability of getting exactly 2 heads in 6 tosses of a fair coin.
Solution:
\[ p = q = \frac{1}{2}, \ n = 6 \]
\[ \binom{6}{2} \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^{6-2} = \frac{6!}{2!(6-2)!} = \frac{15}{64} \]

Properties of the binomial distribution

a. Mean = \( \mu = np \)
b. Variance = \( \sigma^2 = npq \)
c. Standard deviation = \( \sigma = \sqrt{npq} \)
d. Moment coefficient of skewness = \( \frac{q - p}{\sqrt{npq}} \)
e. Moment coefficient of kurtosis = \( 3 + \frac{1}{6pq} \)
f. The distribution is discrete.

Example: Find the mean number of heads in 100 tosses of coin

Solution
This is the expected number of heads in 100 tosses of which is
\[ \mu = np = 100 \left(\frac{1}{2}\right) = 50, \ \text{probability of head} \ p = \frac{1}{2} \]
The standard deviation will be \( \sigma = \sqrt{npq} = \sqrt{100 \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)} = \sqrt{25} = 5 \).

Example: A batch of 30 tablets of soap is taken from a production line, three of which are found to be defective. If an inspector selects two tablets at random from the batch without knowing of the defective tablets, calculate the probability that:

a. Two selected tablets are defective
b. At least one of the selected tablets is defective

Solution
Proportion of defective tablets \( \frac{3}{30} = \frac{1}{10} = 0.10 = p, \ p = 1 - 0.10 = 0.90 \)
a. Probability of 2 defective is \( \binom{30}{2} (0.10)^2 (0.90)^{30-2} = 0.228 \).
b. Probability of at least one defective is \( 1 - \binom{30}{0} (0.1)^0 (0.9)^{30} = 0.958 \).
Example: Find the values of n and p (the no of trials and probability of a "success") for a binomial distribution for which the mean is 7 and variance is $4\frac{2}{3}$.

Solution
For a binomial distribution, the mean = np and the variance is npq.
If np = 7 and npq = 4 2/3

$q = \frac{npq}{np} = 4\frac{2}{3} = 14/3 \times 1/7 = 2/3$

Since $p = 1 - q = 1 - 2/3 = 1/3$, hence,

$n \times 1/3 = 7 \Rightarrow n = 3 \times 7 = 21$.

Poisson Distributions
Poisson distribution is a discrete probability distribution defined for all positive integers in which the probability of exactly x occurrences is given by:

$p(x) = e^{-\lambda} \frac{\lambda^x}{x!}$, for $x = 0, 1, 2, \ldots$

where $\lambda$ is a positive constant called the parameter of the distribution and $e$ is given by the series:

$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \ldots = 2.71828$

Thus, a poisson random variable is a discrete random variable and any non-negative integer is in its range. The poisson distribution provides a probabilistic model for a wide class of phenomena.

Some properties of the Poisson distribution
Mean $\mu = \lambda$
Variance $s^2 = \lambda$
Standard deviation \( s = \sqrt{\lambda} \).
Moment coefficient of skewness \( = 1/\sqrt{\lambda} \).
Moment coefficient of kurtosis \( = 3 + 1/\lambda \).

**Example:** The actuary of a life insurance company has found that the probability of a person having a fatal accident is 0.0002. If the company holds 15000 life insurance policies, what is the probability that the company will pay more than 3 claims next year due to fatal accidents?

**Solution**

\( n = 15,000, \ p = 0.0002 \Rightarrow np = 15000 \times 0.0002 = 3 = m \)

Since \( n > 50 \) and \( np < 5 \), poisson distribution is appropriate here.

The probability that the company will pay \( x \) claims next year is

\[
p(x) = \frac{\lambda^x e^{-\lambda}}{x!}
\]

where \( \lambda = 3 \)

The probability that the company will pay more than 3 claims next year is the probability that it will pay from 4 to 15,000 claims which works out as

\[
1 - p(\text{no claim} + 1 \text{ claim} + 2 \text{ claims} + 3 \text{ claims})
\]

\[
1 - 3^0 e^{-3} + 3^1 e^{-3} + 3^2 e^{-3} + 3^3 e^{-3}
\]

\( 0! \ 1! \ 2! \ 3! \)

\[
1 - [0.04979 + 0.14037 + 0.224055 + 0.224055]
\]

\[
1 - 0.6383 = 0.3617.
\]

**Normal Distribution**

The normal distribution is by far the most widely used continuous distribution and plays a central role in statistical inference. It is also called normal curve or Gaussian distribution defined by the equation:

\[
p(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty.
\]
where $\mu$ is the population mean, $\sigma$ is the population standard, $\pi$ is 3.14159 and $e = 2.71828$. The total area bounded by the curve and the x-axis is one, hence the area under the curve between two ordinates $x = a$ and $x = b$, where $a < b$, represents the probability that $x$ lies between $a$ and $b$, denoted by $\text{pr}(a < x < b)$.

The standard normal distribution is a normal distribution with a mean of 0 and a standard deviation of 1. Normal distributions can be transformed to standard normal distributions by the formula:

$$z = \frac{X - \mu}{\sigma}$$

where $X$ is a score from the original normal distribution, $\mu$ is the mean of the original normal distribution, and $\sigma$ is the standard deviation of original normal distribution. The standard normal distribution is sometimes called the $z$ distribution. A $z$ score always reflects the number of standard deviations above or below the mean a particular score is.

One reason the normal distribution is important is that many psychological and educational variables are distributed approximately normally. Measures of reading ability, introversion, job satisfaction, and memory are among the many psychological variables approximately normally distributed. Although the distributions are only approximately normal, they are usually quite close.

A second reason the normal distribution is so important is that it is easy for mathematical statisticians to work with. This means that many kinds of statistical tests can be derived for normal distributions. Almost all statistical tests discussed in this text assume normal distributions. Fortunately, these tests work very well even if the distribution is only approximately normally distributed. Some tests work well even with very wide deviations from normality.

Finally, if the mean and standard deviation of a normal distribution are known, it is easy to convert back and forth from raw scores to percentiles.

The following features characterize the normal distribution:

a. It is a continuous distribution described by a bell-shaped curve.

b. It is completely determined by its mean which can be any number, positive or negative, and its standard deviation which can be any positive number.
c. It is unimodal and its curve is peaked at the centre and symmetric about a vertical line at the mean.

d. The tails of the curve extend indefinitely in both directions from the centre getting closer and closer to the horizontal axis but never quite touching it.

e. The mean μ determines where the centre of the curve is located and the standard deviation σ determines its flatness.

Some Properties of the Normal Distribution

Mean = μ,  
Variance = σ²

Standard deviation = σ

Moment coefficient of skewness α₃ is 0

Moment coefficient of kurtosis α₄ = 3.

Mean deviation = σ\sqrt{2/\pi} = 0.7979σ

And so the table of standard normal distribution (area from a to z, z positive gives the areas under the standard normal curve from z = 0 to z = a is any number 0.00 0.01 0.02 . . . 3.09.

Example: Find the area between z = 0 and z = 1.73

Thus, the area between 0 and 1.73 is 0.4582 and p(0 < z < 1.73) = 0.4582

Example find p(1.8 < z 2.8)

P(1.8 < z < 2.8) = p(0 < z < 2.8) – p(0 < z < 1.8)
   = 0.4974 – 0.4641 = 0.0333
If a test is normally distributed with a mean of 60 and a standard deviation of 10, what proportion of the scores is above 85? This problem is very similar to figuring out the percentile rank of a person scoring 85. The first step is to figure out the proportion of scores less than or equal to 85. This is done by figuring out how many standard deviations above the mean 85 is. Since 85 is 85-60 = 25 points above the mean and since the standard deviation is 10, a score of 85 is \( \frac{25}{10} = 2.5 \) standard deviations above the mean. Or, in terms of the formula,

\[
z = \frac{X - \mu}{\sigma} = \frac{85 - 60}{10} = 2.5
\]

A z table can be used to calculate that 0.9938 of the scores are less than or equal to a score 2.5 standard deviations above the mean. It follows that only 1-0.9938 = .0062 of the scores are above a score 2.5 standard deviations above the mean. Therefore, only 0.0062 of the scores are above 85.

Suppose you wanted to know the proportion of students receiving scores between 70 and 80. The approach is to figure out the proportion of students scoring below 80 and the proportion below 70. The difference between the two proportions is the proportion scoring between 70 and 80. First, the calculation of the proportion below 80. Since 80 is 20 points above the mean and the standard deviation is 10, 80 is 2 standard deviations above the mean.

\[
z = \frac{X - \mu}{\sigma} = \frac{80 - 60}{10} = 2
\]
A z table can be used to determine that .9772 of the scores are below a score 2 standard deviations above the mean.

To calculate the proportion below 70,

\[
    z = \frac{X - \mu}{\sigma} = \frac{(70 - 60)}{10} = 1
\]

A z table can be used to determine that the proportion of scores less than 1 standard deviation above the mean is 0.8413. So, if 0.1587 of the scores are above 70 and 0.0228 are above 80, then the proportion between 70 and 80 can be computed by subtracting 0.0228 from 0.1587:

\[
    0.1587 - 0.0228 = 0.13590.
\]

Assume a test is normally distributed with a mean of 100 and a standard deviation of 15. What proportion of the scores would be between 85 and 105? The solution to this problem is similar to the solution to the last one. The first step is to calculate the proportion of scores below 85. Next, calculate the proportion of scores below 105. Finally, subtract the first result from the second to find the proportion scoring between 85 and 105.
Begin by calculating the proportion below 85. You can calculate that 85 is one standard deviation below the mean:

\[ z = \frac{X - \mu}{\sigma} = \frac{85 - 100}{15} = -1 \]

Using a z table with the value of -1 for z, the area below -1 (or 85 in terms of the raw scores) is 0.1587.

Doing the same thing for 105,

\[ z = \frac{X - \mu}{\sigma} = \frac{105 - 100}{15} = 0.333 \]

A z table shows that the proportion scoring below 0.333 (105 in raw scores) is .6306. The difference is .6306 - .1587 = .4719. So, .472 of the scores are between 85 and 105.

Probabilities under a general normal curve

If \( x \) is a normal random variable with mean \( \mu \) and variance \( \sigma^2 \) then

\[ p(a < x < b) = p\left( \frac{a - \mu}{\sigma} < z < \frac{b - \mu}{\sigma} \right) \]

In short, the area between a and b under a normal curve with mean \( \mu \) and variance \( \sigma^2 \) is equal to the area under the standard normal curve between \( \frac{a - \mu}{\sigma} \) and \( \frac{b - \mu}{\sigma} \).
The conversion of x values into z values as described here is called standardization. If x is a given value, then its standard score or z-score is determined by

\[ z = \frac{x - \mu}{\sigma} \]

This gives the number of standard deviations that the x value is from the mean \( \mu \).

**Example**

The weight of food packed in a certain containers is a normally distributed random variable with a mean weight of 500g and standard deviation of 5g. Suppose a container is picked at random. Find the probability that it contains

- More than 510g
- Less than 498g
- Between 491 and 498 g

**Solution**

Suppose x denotes the weight of food packed in the container

\[ P(x > 510) = P(z > \frac{510 - 500}{5}) = P(z < 2) = 0.0228 \]

\[ P(x < 498) = P(z < \frac{498 - 500}{5}) = P(z > 0.4) = 0.3446 \]

\[ P(491 < x < 498) = P\left(\frac{491 - 500}{5} < z < \frac{498 - 500}{5}\right) \]

\[ = P(-1.8 < z < -0.4) \]
\[ = P(0.4 < z < 1.8) \]
\[ = P(0 < z < 1.8) - P(0 < z < 0.4) \]
\[ = 0.4641 - 0.1554 \]
\[ = 0.3087. \]
Post-Test

1. The mean annual income of a worker in a certain professional category is N232,000 with a standard deviation of N14,000. Assuming that the distribution of income is approximately normal, what is the probability that two workers picked at random will each earn less than N225,000?

2. A new vaccine was tested on 100 persons to determine its effectiveness. If the claim of the drug company is that a random person who is given the vaccine will develop immunity with probability 0.8, find the probability that:
   a. Less than seventy-four people will develop immunity
   b. Between seventy-four and eighty-five people, inclusive, will develop immunity.
   
   Hint: \( \mu = np \) and \( \sigma = \sqrt{npq} \)

3. If the probability that a child is a son is 0.4. find the probability that in a family of four children there are:
   i. Two sons
   ii. At least two sons
   iii. All girls
   iv. Three girls

4. A machine produces defective items at the rate of 1 in 10. If 100 items produced by this machine are inspected, what is the expected number of defective items? Also find the variance of the number of defective items.

5. Suppose a company received a consignment of 200 boxes of fluorescent tubes, eight tubes to a box. If the probability that a tube is defective is 0.1, how many of the boxes would you expect to have three defective tubes?

6. Give the expression for the probability distribution of a binomial variable \( x \) with parameters \( n \) and \( \theta \) where \( n \) is the total number of independent trials and \( \theta \) the probability of a “success” in a trial.
   a. What does \( x \) represent?
   b. State the mean and standard deviation of the variable.
   c. If the variable \( x \) were to be approximated to a Poisson variate with parameter \( m \), state the condition under which this can
be done. What would then be the approximate probability distribution of \( x \)? Give the mean and standard deviation of the distribution.

7. Give the following characteristics of a binomial distribution \( f(x) \) with parameter \( y \) and number of independent trials \( n \).
   i. Probability function \( f(x) \)
   ii. Mean
   iii. Variance and its standard deviation.

8. The probability that an entering university student will graduate is 0.4. Determine the probability that of 5 students
   i. None will graduate
   ii. One will graduate
   iii. At least one will graduate

9. The probability that goods produced by a factory are defective is 0.002. If 2000 of such goods are produced, what is the probability that:
   i. None will be defective
   ii. Exactly one will be defective
   iii. Less than two will be defective.

10. a. State the mathematical expressions for the probability of the:
   i. normal distribution
   ii. binomial distribution
   iii. Poisson distribution

   b. Write the binomial expansion of
   i. \((p + q)^4\)
   ii. Hence expand \((\frac{5}{6} + \frac{1}{6})^5\)

References


LECTURE ELEVEN

Estimation

Introduction
Statistical inference is the whole process of making an inductive statement about a population from the information obtained from its representative sample. Statistical inference therefore consist the processes of estimation and test of hypothesis (this has been treated in lecture 10 above).

Objectives
The students are expected to master the following after this lecture:
1. definition of estimation and estimate;
2. properties of an estimator;
3. distinguish between point estimate and interval estimate;
4. determine sample size; and
5. confidence intervals for both the mean and proportions.

Pre-Test
1. What is estimation?
2. Define estimate, point estimate and interval estimate
3. State the confidence interval for μ when σ² is known and when σ² is unknown.

CONTENT
What is Estimation?
It is a process by which statistic obtained from a sample are used to estimate the parameters of the population from which the sample was
drawn. Or better still, estimation is the process of calculating from sample data some statistic that is offered as an approximation to the corresponding population parameter of the population for which the sample was selected. The need for estimation arises in practically every statistical application and decision-making especially in both the social sciences and physical sciences. There are two types of estimates – point and interval estimates.

a. **Point Estimate**

A point estimate provides a single numerical value as an assessed value of the parameter under investigation. A point estimate provides a single numerical value as an assessed value of the parameter under investigation. It is a single numerical value used in estimating the corresponding population parameter e.g. \( E(\bar{x}) = \mu \), \( E(p) = p \) and \( E(s) = \sigma / \sqrt{n} \)

The procedure for obtaining a point estimate of a parameter involves the following steps.

i. Pick a random sample of a certain size.

ii. Devise a random variable – statistic.

iii. Compute the value of the statistic from the observed sample data.

The statistic used for the purpose of estimation is called an estimator. It is a random variable which, in some sense, is relevant for estimating the parameter and is simply the form of mathematical expression, that is, a formula. Example: \( \bar{x} = \frac{\sum x}{n} \) is an estimator of \( \mu \). Likewise \( p = \frac{x}{n} \) is an estimator of \( \pi \). Similarly, \( s^2 = \frac{1}{n-1} \sum (x - \bar{x}) \) is an estimator of \( \sigma^2 \).

**Definition:** An estimate is a specific value obtained. Using a particular estimator e.g. \( x = 25 \) implies that 25 is an estimate of \( \mu \), \( P = 0.6 \) implies that 0.6 is an estimate of \( \pi \), similarly \( S^2 = 4 \) implies that 4 is an estimate for \( \sigma^2 \).

A numerical value of the estimator computed from a given set of sample values is called a point estimate of the parameter. Thus, a point estimate is a single number.
Criteria for “Good” Estimators

Many of the basic properties and criteria for characterizing point estimates were formulated and developed by Sir Ronald A. Fisher (1890 – 1962). They include

a. Unbiasedness

An estimator \( \hat{o} \) is an unbiased estimator of a parameter \( o \) if its sampling distribution has mean \( o \), the parameter being estimated. That is, \( E(\hat{o}) = o \). Recall that \( E(\hat{o}) \) means the expected value, or mean, of \( \hat{o} \).

An estimator that is not unbiased is said to be biased. A biased estimator will, on the average, either underestimate or overestimate \( o \).

For example, \( E(\bar{x}) = \mu \), \( E(p) = p \) and \( E(s) = \sigma / \sqrt{n} \)

b. Efficiency

If \( \hat{o}_1 \) and \( \hat{o}_2 \) are two unbiased estimator of \( O \), then \( \hat{o}_1 \) is said to be more efficient than \( \hat{o}_2 \) if the variance of the sampling distribution of \( \hat{o}_1 \) is less than the variance of the sampling distribution of \( \hat{o}_2 \).

c. Consistency

An estimator is consistent for \( \hat{o} \) if it yields values which get closer to the true value of the parameter as sample size increases.

\[
\lim_{n \to \infty} P \left( \left| \bar{x} - \mu \right| > \epsilon \right) = 0 \quad \text{(Law of large numbers)}
\]

d. Sufficiency

Sufficiency is a rather involved concept mathematically. Essentially, what it states is that an estimator should be such that it utilizes all the information all the information contained in the sample for the purpose of estimating a given parameter. No other estimator should provide any more information.

Interval Estimate

This consists of two numerical values computed from sample observation defining an interval which can assert with some degrees of confidence that it includes the population parameter being estimated.
Procedure for Interval Estimation

i. Some degrees of confidence desired of the estimate is stated in advance. This is often expressed in terms of the probability that the interval estimate will cover or include the population parameter. This probability is called the confidence level.

ii. Pick a random sample of a certain size.

iii. On the basis of the observed sample values or observation, two values are computed, these are the endpoints of the interval.

iv. Devise two random variables, which we shall call L and U, thereby getting a random interval (L, U). L and U denote the lower and upper limits of an interval computed from the sample observation and let \( \hat{\theta} \) denotes the population parameter being estimated. If we can assert with probability \( 1 - \alpha \) that the interval \([L, U]\) include the true population parameter \( \theta \), then the interval is said to be \( 100(1 - \alpha)\% \) confidence interval for \( \theta \), therefore

\[
P[L \leq \hat{\theta} \leq U] = 1 - \alpha
\]

The fraction \( 1 - \alpha \) associated with the confidence interval is called the confidence coefficient (confidence level). When the confidence coefficient is expressed as a percentage, we get the level of confidence. The left endpoint of the interval is called the lower confidence limit and the right endpoint, the upper confidence limit.

The method of confidence intervals provides an interval – a range of numerical values – which is believed to contain the unknown parameter.

Confidence interval for mean \( \mu \)

The \( 100(1 - \alpha)\% \) confidence interval for population mean is given as

\[
\text{If: } x \pm \frac{Z_{1-\alpha/2}}{\sigma} \sigma x \text{ if } \sigma^2 \text{ is known} \\
: x \pm \frac{t_{\alpha/2}}{2Sx} Sx \text{ if } \sigma^2 \text{ is unknown}
\]

Where \( x \) is the sample mean based on \( n \) observations.

**Example 1:** For the 90% confidence interval \( 100(1 - \alpha)\% = 90\% \)

\[
\alpha = 0.10, \frac{\alpha}{2} = 0.05, Z_{0.05} = Z_{0.95} = 1.64
\]

204
**Example 2:** For the 95% confidence interval $100(1 - \alpha)\% = 95\%$  
$\alpha = 0.05$, $\alpha/2 = 0.025$, $Z_{1-0.025} = Z_{0.975} = 1.96$  
$FZ(1.96) = 0.975$  
$t_{\alpha/2} = t_{0.05}$, depends on the degree of freedom.  
\[
p \left[ \frac{-1.96 < x - \mu}{\sigma/\sqrt{n}} < 1.96 \right]
\]  
This can be simplified further to give  
\[
\begin{align*}
100(1 - \alpha)\% & = 95\% \\
Z_{1-\alpha/2} & = Z_{0.975} = 1.96 \\
\mu: & 25 \pm (1.96)(3/\sqrt{144}) \\
& 25 \pm (1.96)(0.25) \\
& 25 \pm 0.49 \\
& 24.51 \leq \mu \leq 25.49
\end{align*}
\]

**Example 3:** Given that $n = 144$, $\sigma^2 = 9$, $x = 25$, obtain (1) 95% confidence interval for $\mu$ and (2) 99% confidence interval for $\mu$.

**Solution**

1. $100(1 - \alpha)\% = 95\%$  
   $Z_{1-\alpha/2} = Z_{0.975} = 1.96$  
   $\mu: 25 \pm (1.96)(3/\sqrt{144})$  
   $25 \pm (1.96)(0.25)$  
   $25 \pm 0.49$  
   $24.51 \leq \mu \leq 25.49$

2. $100(1 - \alpha)\% = 99\%$  
   $Z_{1-\alpha/2} = Z_{0.995} = 2.58$  
   $\mu: 25 \pm (2.58)(0.25)$  
   $25 \pm 0.645$  
   $24.355 \leq \mu \leq 25.645$

This is equivalent of saying that we are 95% confident that the population (or true mean) lies between 24.355 and 25.645.

**Example 4:** If a sample of 20 cigarettes is examined for tar content and the mean tar content in this sample is 10.8mgs, taking into account that $\sigma = 0.6$. Compute the confidence interval for mean.
Solution

\[ X \pm 1.96 \frac{\sigma}{\sqrt{n}} = 10.8 \pm 1.96 \frac{0.6}{\sqrt{20}} \]
\[ 10.54 \leq \mu \leq 11.06 \]

We call the interval (10.54, 11.06) a 95% confidence interval for the true mean tar content \( \mu \).

Example 5: A gas station sold a total of 8019 gallons of gas on nine randomly picked days. Suppose the amount sold on a day is normally distributed with a standard deviation \( \sigma \) of 90 gallons. Construct confidence intervals for the true mean of amount sold on a day with the following confidence levels.

a. 98%  

b. 80%

Solution

We are given that \( \sum x = 8019 \), \( n=9 \), \( \sigma = 90 \), therefore \( x = \frac{8019}{9} = 891 \)

a. Because \( \alpha = 0.02 \), \( Z_{\alpha/2} = Z_{0.01} = 2.33 \) from the standard normal table, therefore a 98% confidence interval is given by

\[ \mu : 891 \pm 2.33 \left( \frac{90}{\sqrt{9}} \right) \]
\[ 821.1 < \mu < 960.9 \]

b. \( \alpha = 0.20 \), \( Z_{\alpha/2} = Z_{0.10} = 1.28 \), consequently, with 80% confidence

80 percent confidence is

\[ 891 \pm 1.28 \left( \frac{90}{\sqrt{9}} \right) \], that is,
\[ 852.6 < \mu < 929.4 \]

Example 5: In a time motion study, the sample of a time readings in minutes of an element was taken and the results are as follows:

0.12, 0.14, 0.16, 0.12, 0.12, 0.17, 0.15, 0.14, 0.12, 0.11, 0.12, 0.12, 0.12, 0.15, 0.17, 0.13, 0.14, 0.14.

Within what limit would you expect the actual average time to lie with 95% confidence?
**Solution**

\[ n = 18, \bar{x} = \frac{\sum x}{n} = 0.136, \ S^2 = \frac{1}{n-1}\sum (x - \bar{x})^2 = 0.0185 \]

\[ t_{\alpha/2(17)} \text{df} = t_{0.025}(17) \text{df} = 2.13 \]

\[ \mu: 0.136 \pm (2.13)(0.0185/\sqrt{18}) \]

\[ : 0.136 \pm (2.13)(0.00436) \]

\[ : 0.136 \pm 0.0092868 \]

\[ 0.1267 \leq \mu \leq 0.1453 \]

**Determination of Sample Size**

From the knowledge gained from the confidence limits, it is written as

\[ \mu: x + Z_{1-\alpha/2}(\sigma/\sqrt{n}) \] which can be written in the form

\[ \mu: x + \text{Error term} \]

\[ \therefore \text{Error term} = \varepsilon = Z_{1-\alpha/2}(\sigma/\sqrt{n}) \]

Making \( n \) the subject

\[ \Rightarrow n = \left( Z_{1-\alpha/2}(\sigma/\sqrt{n}) \right)^2 \]

\[ \varepsilon \]

**Example 6:** A machine is designed to produce rubber gasket with a mean thickness of 0.125cm, the standard deviation is 0.01cm determine its sample size.

**Solution**

\[ 1 - \alpha = 1 - 0.95 = 0.05, \ \alpha/2 = 0.025, \ \sigma = 0.01, \ \varepsilon = 0.05 \]

\[ \therefore n = \left( Z_{1-\alpha/2}(\sigma/\sqrt{n}) \right)^2 \]

\[ \varepsilon \]

\[ n = \frac{(1.96)(0.01)}{0.05} \]

\[ = (19.6)^2 \]

\[ = 384.16 \]

\[ n = 384 \]

The sample size should be at least 384.
Summary
At times, we may have no information or the information may be scanty about some characteristics of the population, especially, the values of the parameters involved in the distribution, and it is required to obtain estimates of these parameters. This is the problem of Estimation. All these attempts are geared towards looking for acceptable intervals that would contain sample statistics that would be consistent, efficient, unbiased and sufficient to account for the whole population.

Post-Test
1. For the given level of confidence in each case, determine $\alpha$
   i. 96%
   ii. 92%
   iii. 84%
   iv. 97%
2. Determine the indicated value for the given level of confidence
   a. $Z_{\alpha/2}$; 90%
   b. $Z_{\alpha/2}$; 93.84%
   c. $Z_{\alpha/2}$; 97.5%
   d. $-Z_{\alpha/2}$; 91.64%
   e. $-Z_{\alpha/2}$; 97.42%
3. From past records, the seasonal rainfall in a country when observed over sixteen randomly picked years yielded a mean rainfall of 20.8 inches. If it can be assumed from past experience that rainfall during a season is normally distributed with $\sigma = 2.8$ inches. Construct confidence intervals for the true mean rainfall $\mu$ with the following confidence levels:
   i. 90%
   ii. 95%
   iii. 98%
   iv. $t_{4,0.05}$
   v. $t_{21,0.05}$
4. An electrical firm that manufactures a certain type of bulb wants to estimate its mean life. Assuming that the standard error is known to be 40 hours, find how many bulbs should be tested so as to be: 
   a. 95% confident that the estimate $x$-bar will not differ from the true mean life $\mu$ by more than 10 hours?
   b. 98% confident to accomplish accuracy in (a)
5. When ten vehicles were observed at random for their speeds (in mph) on a freeway, the following data were obtained: 61, 53, 57, 55, 62, 58, 60, 54, 63, 60. Suppose that from experience it may be assumed that vehicle speed is normally distributed with $\sigma = 3.2$ mph. Construct a 90% confidence interval for $\mu$, the true mean speed.
6. In each of the following, obtain point estimates of the population mean, variance, and standard deviation.
   a. $n = 12$, $\sum x = 1080$, $\sum (x - \bar{m})^2 = 6875$
   b. $n = 10$, $\sum x = 185$, $s^2 = 1.764$

References

Arinola, J.A: Handout on Business Statistics written for Delta State University, Abraka, (Ibadan Study Centre).
ICAN ATS Statistics Handout
LECTURE TWELVE

Tests of Hypotheses

Introduction
A hypothesis is an empirically-testable statement about a relationship involving two or more variables. Examples of hypotheses from the social sciences include:

- Investors seek low-risk investments in economic downturns.
- Students’ feelings of connectedness to school are an essential element of their academic success.

Each of these specifies a relationship that may or may not exist under particular conditions. They are testable statements about relationships between different factors. But why bother with forming a hypothesis as part of the research process? All these will be seen in this lecture.

Objectives
At the end of this lecture, readers are to have mastered the following:
1. the meaning of hypothesis;
2. the types of hypotheses in the social sciences;
3. basic tests in hypothesis testing; and
4. real life hypothesis formulations.

Pre-Test
1. Define the following with respect to hypothesis testing:
   i. composite hypothesis
   ii. alternate hypothesis
2. Explain a one-tailed and two-tailed test with the aid of appropriate diagrams.

3. A random sample of 200 commercial banks show a mean profit of N120m with a standard deviation of N25m. Find the probability that the mean profit of commercial banks is at least:
   a. N125m
   b. N110
   Test at 1% significance level.

4. The manageress of Christie supermarket stated that the average number of snacks sold per day was 250 with a variance of 225. A worker in the supermarket wants to test the accuracy of the claims. The worker took samples of the sales for 25 days and the average number of snacks sold per day was 243. Using 0.05 level of significance, should the claim of the manageress be accepted?

5. Explain the following statistical terms as they relate to hypothesis testing.
   a. Critical region
   b. Type I error
   c. Level of significance

6. State clearly, the null and alternative hypothesis of test of two means for one-tailed test and two-tailed test.

7. It is claimed that the Distance Learning graduates employed by Procter and Gamble commit average of 13 calculation errors per week. A random check conducted on ten of them resulted in the following average number of errors per week: 15, 19, 22, 17, 13, 16, 12, 24, 20, 14. Test the hypothesis at 95% confidence level that the claim is an understatement.

8. An Economist in 2001 compared the wages of Federal Government and State Government employees. He discovered that the mean monthly wage of Federal Government employees is N1,000 more than that of State Government employees. He repeated the study in 2006, since he wonders whether or not there
is still the same difference in their pay. He finds that the mean monthly wage of 100 Federal Government employees, chosen at random, is ₦17,000 with a standard deviation of ₦2000, while a random sample of 81 state Government employees has a mean wage of ₦15,500 with a standard deviation of ₦1,000. At a 0.05 level of significant, should the economist conclude that the difference in wages of Federal Government and State Government employees is unchanged?

9. A stockbroker claims that he can predict with 80% accuracy whether a stock market value will rise or fall during the coming month. As a test, he predicts the outcome of 40 stocks and is correct in 28 of the predictions. Does this evidence support the stockbroker’s claim?

10. A sample survey conducted among the inhabitant of Ibadan recently shows that 627 of 800 persons interviewed prefer to live in the suburb areas of Ibadan. Test the hypothesis that the true proportion of persons who prefer to live in the suburb is 0.75.

11. In a big city, 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers? State the hypothesis clearly.

CONTENT

What is a Hypothesis?

“A statistical hypothesis is a statement, assertion or claim about the nature of a population.” It is a basic object of any experimental inquiry.

A statistical hypothesis is a method of making statistical decisions using experimental data. It is sometimes called confirmatory data analysis in contrast to exploratory data analysis.

Many a time, we strongly believe some results to be true. But after taking a sample, we notice that one sample data does not wholly support the result. The difference is due to:

i. the original belief being wrong and

ii. the sample being slightly one sided.

Tests are therefore, needed to distinguish between the two possibilities. These tests tell about the likely possibilities and reveal whether or not the difference can be due to chance elements. If the
difference is not due to chance elements it is significant and therefore, these tests are called tests of significance. The whole procedure is known as Testing of Hypothesis.

A hypothesis is a statement supposed to be true till it is proved false. It may be based on previous experience or may be derived theoretically. First a statistician or the investigator forms a research hypothesis that an exception is to be tested. Then he derives a statement which is opposite the research hypothesis (noting as H_o). The approach here is to set up an assumption that there is no contradiction between the believed result and the sample result and that the difference therefore can be ascribed solely to chance. Such a hypothesis is called a null hypothesis (H_o). It is the null hypothesis that is actually tested, not the research hypothesis. The object of the test is to see whether the null hypothesis should be rejected or accepted.

If the null hypothesis is rejected, that is taken as evidence in favor of the research hypothesis which is called the alternative hypothesis (denoted by H_a). In usual practice we do not say that the research hypothesis has been "proved" only that it has been supported.

For example, if it is assumed that the mean of the weights of the population of a college is 110 lb, then the null hypothesis will be the mean of the population that is 110 lbs, i.e. H_o : m = 110 lbs (Null hypothesis).

In terms of alternative hypothesis
i. H_a : m $\neq$ 110 Kg
ii. H_a : m > 110 Kg
iii. H_a : m < 110 Kg.

In the framework of statistical investigation, very often a hypothesis takes the form of stipulating the values of the unknown parameters of the population being studied. For example, consider the following.

The statement that the average height of the student in the class is 1.6 metres, the hypothesis is that $\mu = 1.6$.

An essential component of the scientific process is the formulation and evaluation of hypotheses. In seeking to learn more about the social world, social scientists ask many different kinds of questions about relationships between factors of social life. How do investors change their behaviour when market conditions change? What role did political and social factors play in the corruptions in the country? Do feelings of
connectedness influence students' performance in school? To address these questions, social scientists form hypotheses which they then evaluate using some form of data.

You may be familiar with examples of hypotheses and hypothesis testing from the natural sciences, perhaps through schoolwork or participation in a school science fair. You may have evaluated hypotheses such as:

- The combination of certain chemical compounds yields heat energy.
- Plants' growth is enhanced through exposure to ultraviolet light.
- When a moving object collides with another object, the total kinetic energy of the two objects does not change.

Typically, hypotheses such as these are generated from some theory or theoretical perspective, and then evaluated using data collected through some laboratory procedures. Research in the social sciences works similarly (though, often outside the laboratory). This module is designed to introduce you to hypothesis in the social sciences.

**Definition**: The hypothesis which is equivalent to a complete specification of the distribution of a random variable is called a simple hypothesis. Otherwise it is called a composite hypothesis.

**Note**: A distribution is completely specified by giving the functional form and all the parameters of the distribution. E.g. If the is known to be normal and have a variance \( \sigma^2 \), the hypothesis that \( \mu = \mu_0 \) is a simple hypothesis since the mean and the variance both specifies a normal distribution population. If however, variance is unknown, the hypothesis is said to be a composite hypothesis.

**Definition**: “A hypothesis being tested for the purpose of possible rejection is called a null hypothesis.” It is a common convention to designate this hypothesis as \( H_0 \). A null hypothesis is the hypothesis of no change, that is, no difference. It is often a negation of the statement or claim of particle of interest actually being tested.
\begin{align*}
  \text{Ho: } p &= \frac{1}{2} & \text{Ho: } \mu &= \mu_0 & \text{Ho: } \sigma_1^2 &= \sigma_2^2 \\
  \text{Ho: } \pi &= \pi_0 & \text{Ho: } \mu &\geq 700.
\end{align*}

Generally speaking, in social science research, we establish two competing hypotheses that we then evaluate in light of some empirical data. These hypotheses are referred to as the null hypothesis and the alternative hypothesis. The primary purpose of hypothesis testing is to examine the likelihood of the null hypothesis with data.

**Definition:** The hypothesis against which the null hypothesis is tested is called the alternative hypothesis. We denote it as $H_1$ or $H_A$. This is the hypothesis that is accepted when the null hypothesis is rejected.

The values of the parameters stated under the alternative hypothesis are outside the region stated under the null hypothesis. That is, there is no overlap between the set of parameter values stipulated under the null hypothesis and those stipulated under the alternative hypothesis.

- $H_1: \mu \neq \mu_0$
- $H_1: \mu < \mu_0$
- $H_1: \mu > \mu_0$

**Choice of null hypothesis ($H_0$) and alternative hypothesis ($H_1$)**

In a given problem, which hypothesis constitutes the null hypothesis and which one is the alternative hypothesis is an important question and one that has to do with the logic of a statistical test procedure. In formulating tests of hypotheses, the essential idea is that of proof by contradiction. For example, suppose a teacher strongly suspects that students coming from families at higher income level perform better than students of low income earners. This is what the teacher believes she can establish and so the burden of proof is on her. Under $H_0$ we will say that this is not the case and state:

- $H_0$: students at higher economic level do not perform better
- $H_1$: students at a higher economic level do perform better
Thus it is the claim of the researcher that it is stated as the alternative hypothesis. For this reason, the alternative hypothesis is sometimes referred to as the research hypothesis. The null hypothesis asserts, in essence, that there is no merit to the researcher’s claim.

Formulation of $H_0$ and $H_1$

When we wish to establish a statement about the population with evidence obtained from the sample, the negation of the statement is what we take as the null hypothesis $H_0$. The statement itself constitutes the alternative hypothesis $H_1$.

**Definition**: A test of hypothesis is a rule or procedure that leads to a decision to accept or reject the hypothesis when the experimental sample values are obtained. This rule, formulated before drawing the sample, is often referred to as a decision rule.

The alternative hypothesis may be either directional or non-directional when $H_1$ only assert that the population parameter is different from the one hypothesized, this is referred to as a non-directional or two-tailed hypothesis. For example, if the alternative hypothesis is $H_1: \mu \neq \mu_0$, this is a two-tailed test. If alternative hypothesis is a one-directional test, then the alternative hypothesis can be either

- $H_1: \mu < \mu_0$ if the suspicion is that the population mean is less than $\mu_0$.
- $H_1: \mu > \mu_0$ if the suspicion is that the population mean exceeds $\mu_0$.

If a researcher has no idea whether a population mean is greater or less than some hypothesized value, then a two directional test is most commonly used. All the researcher may need to know is whether the sample mean supports the hypothesis or not, so he or she uses a two-directional test, with the alternative hypothesis being that $\mu$ is not equal to the value specified in the null hypothesis. The critical region, or region of rejection, is then in the two tails of the distribution. For such test, the critical region is on both sides of the sampling distribution of the test statistic.
If the question gives some hint that the population mean may exceed the hypothesized value, then the alternative hypothesis is $H_I: \mu > \mu_0$ and the region of rejection is in the extreme right tail of the distribution. Similarly, when the question gives some hint that the population mean may fall short of the hypothesized value, then the alternative hypothesis is $H_I: \mu < \mu_0$ and the region of rejection is in the extreme left tail of the distribution.

**Definition**

A test statistic is a numerical value computed from a sample observation on the assumption that the population follow a particular distribution (binomial, normal, poisson, $\chi^2$, etc.) used to decide whether or not to reject the null hypothesis.

In the case of hypothesis concerning the population mean $\mu$, the test statistic is

$$Z = \frac{x - \mu}{\sigma/\sqrt{n}} \sim N(0,1),$$

$$t = \frac{x - \mu}{s/\sqrt{n}} \sim t_{n-1}.$$
Definition

The critical region of a test is that region such the null hypothesis is rejected if the test statistic computed from a particular sample of a given size takes on a value that lies within this region.

\[ z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \]

Now \( z \) i.e. the relative deviation of sample is a standard normal variate from which it is extremely useful to draw statistical inference.

We know that the standard normal variate is our \( z \)-score. 95% area under the curve lies between -1.96 and +1.96, 99% lies between -2.58 and +2.58 and 99.73% lies between -3 and +3. In other words, only 5% area under the normal curve lies beyond \( \pm 1.96 \), 1%, beyond \( \pm 2.58 \) and 0.27% lies beyond \( \pm 3 \). These areas also indicate the probabilities that the \( z \)-score exceeds these values i.e. the probability that \( z \)-score will exceed 1.96 numerically is 0.05.

- \( P (z < -1.96 \text{ or } z > 1.96) = 0.05 \)
- \( P (|z | > 1.96) = 0.05 \)

This means that the probability that \( z \)-score will lie in the shaded region is 0.05 which is very small.

Similarly \( P (|z | > 2.58) = 0.01 \) and \( P (|z | > 3) = 0.0027 \).

Return to our \( z \)-score formula \( z_1 = \frac{\bar{x}_1 - \mu_0}{\sigma/\sqrt{n}} \) where mean of sample is for which \( z \)-score is \( z_1 \).

Now 95 out of 100 values of \( z_1 \) will lie between -1.96 and +1.96. For \( z_1 > 1.96 \text{ or } z_1 < 1.96 \), a rare event has taken place because \( P (|z_1 | > 1.96) = 0.05 \) which is very small. The \( z \)-score \( z_1 \) of \( x_1 \) from \( \mu_0 \) is so significant that it can not due to sample fluctuations alone.

If \( P (|z_1 | > 2.58) = 0.01 \), then very unusual event has taken place because this probability is extremely small (remote). Again we can firmly say that the deviation of \( x_1 \) from \( \mu_0 \) (i.e. \( z \)-score) is significant and it is not due to the sample fluctuations alone.

From this discussion we come to a strong conclusion that the levels marked by the probabilities 0.05 or 0.01 which determine the significance of an event are called **Levels Of Significance** and are always expressed in
percentages as 5% level of significance or 1% level of significance. The corresponding regions are called **Critical Regions**.

The limits within which we expect $z$-score lies with specified probabilities are called 'Confidence limits'. Thus $P(|z| > 1.96) = 0.05$ and the bounding values are $\pm 1.96$ are the confidence limits or fiducial limits. It means we are confident that 95 cases out of 100, the sample mean $x$ will be such that $z$ lies between $-1.96$ and $+1.96$.

Main objects of sampling theory are

i. Estimation

ii. Testing a hypothesis.

**Setting up levels of significance**: Once the null hypothesis is set up, the next job is to set the limits within which we expect (the null hypothesis) $m$ lies. The idea behind it is to ensure that the difference between the sample value and the hypothesis should arise due to sampling fluctuations alone. If this difference does not exceed this limit then the sample supports the null hypothesis and the sample is accepted. If it exceeds this limit the sample does not support the hypothesis and it is rejected.

Now fixing the limits totally depends upon the accuracy desired. Generally the limits are fixed such that the probability that the difference will exceeds the limits is 0.05 or 0.01. These levels are known as the 'levels of significance' and are expressed as 5% or 1% levels of significance. Rejection of null hypothesis does not mean that the hypothesis is disproved.

It simply means that the sample values do not support the hypothesis. Also, acceptance does not mean that the hypothesis is proved. It means simply it is being supported.
**Confidence limits**: The limits (or range) within which the hypothesis should lie with specified probabilities are called the confidence limits or fiducial limits. It is customary to take these limits as 5% or 1% levels of significance. If sample values lies between the confidence limits, the hypothesis is accepted; if it does not, the hypothesis is rejected at the specified level of significance.

Type I Error and Type II Error

**Errors in Testing Of Hypothesis**

In testing any hypothesis, we get only two results: either we accept or we reject it. We do not know whether it is true or false. Hence four possibilities may arise.

1. The hypothesis is true but test rejects it (Type I error)
2. The hypothesis is false but test accepts it (Type II error)
3. The hypothesis is true and test accepts it (correct decision)
4. The hypothesis is false and test rejects it (correct decision)

In a statistical hypothesis testing experiment, a Type I error is committed when the null hypothesis is rejected though it is true. In terms of probability, Type I error is denoted by \( \alpha \) (alpha) where

\[
\alpha = \text{probability of type I error} = \text{probability (rejecting } H_0 / H_1 \text{ is true)}.
\]

A Type II error is committed by not rejecting (i.e. accepting) the null hypothesis, when it is false. The probability of Type II error is denoted by \( \beta \) (beta) where

\[
\beta = \text{probability of Type - II error} = \text{probability (accepting } H_0 / H_1 \text{ is false)}
\]

Now consider that the difference between two population means is actually zero i.e. \( H_0 : \mu_1 - \mu_2 = 0 \). But our test says that the difference is significant. Here we make a Type I error.

On the other hand, suppose there is true difference between the population means i.e. \( H_0 : \mu_1 - \mu_2 = 0 \), but our test says the difference is not significant. Here we commit a Type II error.
Type I error is the error committed by rejecting null hypothesis $H_0$ when it is true. The probability of committing type I error is called the ‘level of significance’. Or the size of the test and it is denoted by $\alpha$ where $\alpha = p(\text{rejecting } H_0 / H_0 \text{ is true}).$

Type II error is the error committed by accepting $H_0$ when it is false. The probability of committing type II error is called $\beta = p(\text{accept } H_0 / H_0 \text{ is false}).$ It is called an acceptance error.

In any hypothesis testing problem, because we take action based on incomplete information, there is a built-in danger of an erroneous decision. A statistical test procedure based on sample data will lead to precisely one of the following four situations as given in the table below. Two of these situations will entail correct decisions and the other two, incorrect decisions.

Four possibilities based on the decision taken and the truth or falsity of $H_0$.

<table>
<thead>
<tr>
<th>Test procedure conclusion</th>
<th>True state of nature</th>
<th>Ho is true</th>
<th>Ho is false (H$_1$ is true)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accept Ho</td>
<td>Correct decision; possibility is $1 - \alpha$</td>
<td>Incorrect decision; Type II error; probability is $\alpha$</td>
<td></td>
</tr>
<tr>
<td>Reject Ho (Accept $H_1$)</td>
<td>Incorrect decision; Type I error; probability is $\beta$</td>
<td>Correct decision, probability is $1 - \beta$</td>
<td></td>
</tr>
</tbody>
</table>

This table can be drawn in another way to reflect the same idea.

<table>
<thead>
<tr>
<th>Hypothesis ($H_0$)</th>
<th>Decision Accept ($H_0$)</th>
<th>Reject ($H_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct decision</td>
<td>Type I error</td>
</tr>
<tr>
<td>False</td>
<td>Type II error</td>
<td>Correct decision</td>
</tr>
</tbody>
</table>

Now suppose a manufacturer produces some type of articles of good quality. A purchaser by chance selects a sample randomly. It so happens that the sample contains many defective articles and it leads the purchaser
to reject the whole product. Now, the manufacturer suffers a loss even though he has produced a good article of quality. Therefore, this Type I error is called "producers risk".

On the other hand, if we accept the entire lot on the basis of a sample and the lot is not really good, the consumers are put in loss. Therefore, this Type II error is called the "consumers risk".

In practical situations, still other aspects are considered while accepting or rejecting a lot. The risks involved for both producer and consumer are compared. Then Type I and Type II errors are fixed; and a decision is reached.

**Power of a Hypothesis Test**

The measure of how well the test of hypothesis is working is called the 'power of the test'. In hypothesis testing both $\alpha$ (the probability of Type-I error) and $\beta$ (the probability of Type II error) should be small. Type I error occurs when we reject $H_0$ which is true and $\alpha$ (the significance level of the test) is the probability of making type - I error. Once the significance level is fixed, nothing can be done about $\alpha$. Type II error occurs when we accept $H_0$ which is false. The probability of type II error is $\beta$. The smaller the value of $\beta$, the better is the test. Alternately $(1 - \alpha)$ i.e. the probability of rejecting $H_0$ when it is false should be as large as possible.

Rejecting $H_0$ when it is false is exactly what a good test must do. A high value of $(1 - \alpha)$, mostly close to 1, means that the test is working in high gear. A low value of $(1 - \alpha)$, mostly close 0, means that the test is working poorly.

Thus $(1 - \alpha)$ is the measure of the power of the test. If we plot the values of $(1 - \alpha)$ for each value of $\beta$, for which $H_\alpha$ is true, the resulting curve is known as a power curve. You can see from the figure given below that the power is simply $(1 - \alpha)$. In testing of a hypothesis high power is desirable. As we know that the Type I error or $\alpha$ is usually set in advance by a researcher but the Type II error or $\beta$ for a given test is harder to know as it requires estimation of the distribution of $H_\alpha$, which is unknown in most of the cases. So like $\beta$, the power can be difficult to estimate accurately, but increasing the sample size $n$, we can increase the power of the test.
Procedure for Hypothesis Testing

1. State the underlined assumptions
2. State the null and alternative hypothesis
3. State the level of significance $\alpha$
4. Determine the appropriate test statistic and compute the observed value of the test.
5. Determine the critical region
6. State the decision rule and
7. State your conclusions.

Some Specific Test

1(a) Testing for $\mu$ when $\sigma^2$ is known, the population is normal or $n > 30$ for non normal population.

<table>
<thead>
<tr>
<th>Ho</th>
<th>H1</th>
<th>Reject Ho, if</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu = \mu_0$</td>
<td>$\mu \neq \mu_0$</td>
<td>$</td>
</tr>
<tr>
<td>$\mu = \mu_0$</td>
<td>$\mu &gt; \mu_0$</td>
<td>$Z &gt; Z_{1-\alpha}$</td>
</tr>
<tr>
<td>$\mu = \mu_0$</td>
<td>$\mu &lt; \mu_0$</td>
<td>$Z &lt; -Z_{1-\alpha}$</td>
</tr>
</tbody>
</table>

\[
Z = \frac{x - \mu}{\sigma/\sqrt{n}} \sim N(0,1)
\]
Sampling of Variables

The sample mean \( \bar{x} \) is normally distributed with mean \( \mu \) and standard deviation \( \sigma \), where \( \mu \) is the mean of the population and \( \sigma \) is the standard deviation of the population. This result will help us in testing a hypothesis about the population mean.

1. Testing the hypothesis that the population mean \( \mu = \mu_0 \) which is a specified value.

Now \( \sigma = \text{S.D. of the population is known.} \)

In notations:

\[
|z| = \left| \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \right|
\]

- If \( |z| < 1.96 \), the difference is not significant at 5% level and \( H_0 \) is accepted, otherwise rejected.
- If \( |z| < 2.58 \), the difference is not significant at 1% level and \( H_0 \) is accepted, otherwise rejected.

**Note:** We have assumed that \( \sigma \) is known. If however \( \sigma \) is not known, we take \( s \) to be equal to the S.D. of the sample.

**Example 1:** Ayarjo enterprises supplies bags of rice to Ajombadi hotel. The specification is that each bag should weigh 50kg. A random sample of 16 bags and the mean weight was found to be 49.2kg with a standard deviation of 2kg. At 5% level of significance, would you say that Ayarjo is satisfying the specification?

**Solution**

- \( H_0: \mu = 50 \text{kg} \)
- \( H_1: \mu \neq 50 \text{kg} \)

\[
X = 49.2\text{kg} \\
S = 2\text{kg}, n = 16 \\
\sigma x = s / \sqrt{n} = 2 / \sqrt{16} = 0.5
\]
Test statistic
\[ Z = \frac{x - \mu}{\sigma_x} = \frac{49.2 - 50}{0.5} = 0.8 \]
\[ = \frac{0.8}{0.5} = 1.6 \]

\[ Z_{\text{tabulated}} = 1.96 \]

Since 1.6 is less than 1.96, Ho is accepted, that is, Ayarjo is satisfying the specification.

**Example 2**: From a normal population, the variance 16 unit, a random sample size of 36 items gave a mean of 84 units at 5% level of significance. Test the hypotheses:

a. \( H_0 : \mu = 80 \)

b. \( H_1 : \mu > 80 \)

**Solution**

a. \( H_0: \mu = 80 \)

\( H_1: \mu \neq 80 \)

\( \alpha = 0.05 \)

\[ Z = \frac{84 - 80}{4/\sqrt{36}} = \frac{4/4/6}{6} = 6 \]

\[ Z_0 = \alpha /2 = Z_{0.975} = 1.96 \]

**Conclusion**: Since 6 > 1.96, we reject Ho. The data support the claim that \( \mu \neq 80 \). There is evidence to show that \( \mu \neq 80 \).
b. Ho: \( \mu = 80 \)  
H1: \( \mu > 80 \)  
\( \alpha = 0.05 \)  

\[ Z = \frac{84 - 80}{4\sqrt{36}} = 6 \]

**Conclusion:** Since 6 > 1.645 Reject Ho. The claim that \( \mu > 80 \) is justified by the available data.

**Example** A random sample of 400 male students has average weight of 55 kg. Can we say that the sample comes from a population with mean 58 kg. with a variance of 9 kg. ?

**Solution:** The null hypothesis \( H_0 \) is that the sample comes from the given population.
In notations : \( H_0 : \mu = 58 \) kg. and \( H_1 : \mu \neq 58 \) kg.

\[
|z| = \left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right|
\]

Insert \( \bar{x} = \text{sample mean} = 55 \) kg.
\( \mu = \text{population mean} = 58 \) kg. \( n = 400 \) and \( \sigma = \text{population S.D.} = 3 \)

Therefore

\[
|z| = \left| \frac{55 - 58}{2/\sqrt{400}} \right| = 20 \times 2.58
\]

This value is highly significant. We will reject \( H_0 \) on the basis of this sample. The sample therefore, is not likely to be from the given population.

**Example** A random sample of 400 tins of vegetable oil and labeled "5 kg. net weight" has a mean net weight of 4.98 kg. with standard deviation of 0.22 kg. Do we reject the hypothesis of net weight of 5 kg. per tin on the basis of this sample at 1% level of significance ?
**Solution:** The null hypothesis $H_0$ is that the net weight of each tin is 5 kg.

In notations $H_0: \bar{\theta} = 5$ kg.

Inserting, $\bar{x} = 4.98$ kg., $\theta = 5$ kg. $\theta = 0.22$ kg. and $n = 400$

$$|z| = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \text{ we get}$$

$$|z| = \frac{4.98 - 5}{0.22/\sqrt{400}} = 1.8 < 2.58$$

Hence $H_0$ is accepted at 1% level of significance.

**Sampling of Attributes**

Characteristics like religion, language, locality etc. cannot be measured in numbers as they are attributes. We can only say whether a particular attribute is present or absent in an individual. The sampling of attributes means drawing a sample from the population, of which every member possesses, a particular attribute or does not have it. In many cases, we are interested in knowing how many in the population possess the attribute. For instance, Brooke wants to know how many from his area are smokers i.e. he is interested in the number or proportion of smokers in his area. Now we shall see how to test a hypothesis about the number or proportion of the individuals of the population possessing an attribute.

Here is the hypothesis setting for proportions

2. **Test for $\pi$**

<table>
<thead>
<tr>
<th>$H_0$</th>
<th>$H_1$</th>
<th>Reject $H_0$, if</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\pi = \pi_0$</td>
<td>$\pi \neq \pi_0$</td>
<td>$</td>
</tr>
<tr>
<td>$\pi = \pi_0$</td>
<td>$\pi &gt; \pi_0$</td>
<td>$Z &gt; Z_{1-a}$</td>
</tr>
<tr>
<td>$\pi = \pi_0$</td>
<td>$\pi &lt; \pi_0$</td>
<td>$Z &lt; -Z_{1-a}$</td>
</tr>
</tbody>
</table>

$$Z = \frac{p - \pi}{\sigma p} = \frac{p - \pi}{sp} \sim N(0,1)$$
Testing the hypothesis that population proportion $p = P_o$

In notations: $H_0 : \text{population proportion} = P_o$

Now $P' = \text{sample proportion}$ is normally distributed with mean $P_o$ and standard deviation $\sqrt{\frac{P_o \cdot q_o}{n}}$.

Hence $z = \frac{P' - P_o}{\sqrt{P_o \cdot q_o / n}}$

i. If $|z| > 1.96$ a rare event has taken place.
   If $P(|z| > 1.96) = 0.05$ we reject the hypothesis $P = P_o$ at 5% level of significance.

ii. If $|z| > 1.96$ a usual event has happened and accept the hypothesis that $P = P_o$ at 5% level of significance.

iii. If $|z| > 2.58$ a rare (unusual) event has taken place. Now $P(|z| > 2.58) = 0.01$. We then reject the hypothesis $P = P_o$ at 1% level of significance.

iv. If $|z| > 2.58$, an unusual event occurred and we accept the hypothesis at 1% level of significance.

\[\text{Example}\] A sample of 600 persons selected at random from a Agbowo, near University of Ibadan shows that there are 53% smokers. Is there any reason to doubt the hypothesis that smokers and non-smokers are equal in number in the city?

\[\text{Solution}\] The null hypothesis is that the smokers and non-smokers are equal in numbers.

$H_0 : P = P_o$ and $H_a : P \neq P_o$

where $P_o = 0.5 \quad q_o = 1 - P_o = 0.5$

Also the sample proportion is $P' = 53\% = 0.53$ and $n = 600$.

\[\therefore |z| = \frac{|P' - P_o|}{\sqrt{P_o \cdot q_o / n}} = \frac{0.53 - 0.5}{\sqrt{0.5 \times 0.5 / 600}} = 1.5 < 1.96\]
Hence the hypothesis $H_0$ is accepted at 5% level of significance. i.e. we can undoubtedly say that both smokers and non-smokers are equal in numbers in that city.

**Example** In a sample of 900 stock-holders of companies, 400 stated that their major aim in holding stocks is capital appreciation. What is the 90% confidence range within which lies the population proportion of stock-holders who hold stocks for capital appreciation?

**Solution:** The 90% confidence range is given by $P' \pm 1.64 \text{ S.E.}$ where $\text{S.E.} = \sqrt{\frac{P'q'}{n}}$

$\text{Now } P' = \frac{400}{900} = \frac{4}{9} \text{ and } q' = 1 - p = \frac{5}{9}$

Hence the range is given by $\frac{4}{9} \pm 1.64 \sqrt{\frac{0.4444 \times 0.5556}{900}}$

$= \frac{4}{9} \pm 1.64 \times \sqrt{\frac{20}{270}} = 0.4444 \pm 0.0272$

Hence the range is 0.4444 + 0.0272 = 0.4716 and 0.4444 - 0.0272 = 1.4172 i.e. (0.4172 to 0.4716). Note that when we do not know $p_0$ and $q_0$ then we can use $p'$ and $q'$ in their places respectively.

**Example** Out of consignment of 10000 tennis balls, 400 were selected randomly and examined. It was found that 20 of these were defective. How many defective balls you can reasonably expect to have in the whole consignment at 95% confidence level?

**Solution:** $P' = \frac{20}{400} = 0.05$ and $q' = 1 - p = 0.95$

The limits are (at 95% confidence level) = $p' \pm 1.96 \text{ S.E.}$

where $\text{S.E.} = \sqrt{\frac{p'q'}{n}} = \sqrt{\frac{0.05 \times 0.95}{400}} = 0.01083$
Therefore, $0.05 \div 1.96 \div 0.01089$ i.e. $0.05 \div 0.0213$

Therefore $p'$ lies between $0.0713$ to $0.0287$.
The expected number lies between
$NP' = 10000 \div 0.0713 = 7130$ and
$NP' = 10000 \div 0.0287 = 2870$.

**Example 4**: A party stalwart claims that he has support of 55% of the voters in his constituency. What will the party executive committee conclude if out of a random sample of 500 registered voters only 245 expressed their preference for him?

**Solution**

$H_0: \pi = 0.55$

$H_1: \pi < 0.55$

$\alpha = 0.05$

$p = 245/500 = 0.49$

$sp = \sqrt{(0.49)(0.51)} = 0.022$, that is $sp = \sqrt{\frac{np(1-p)}{n}}$

$Z = 0.49 - 0.55 = -2.73$

$0.022$

$-Z_{1-\alpha} = -Z_{0.95} = -1.645$

**Conclusion**: Since $-2.73 < -1.645$

Reject $H_0$. Therefore the claim by the party stalwart for the support of 55% of the voters in his constituency is not justified by the available data.

**Example** A group of 200 students have the mean height of 154 cms. Another group of 300 students have the mean height of 152 cms. Can these be from the same population with S.D. of 5 cms?

**Solution**: $H_0 : \mu_1 = \mu_2$, the samples are from the same population.
$H_a : \mu_1 \neq \mu_2$, here $\bar{x}_1 = 154$ cms, $\bar{x}_2 = 152$ cms, $s = 5$ cms, $n_1 = 200$ and $n_2 = 300$.

$$|z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| = \left| \frac{154 - 152}{\sqrt{\frac{1}{200} + \frac{1}{300}}} \right| = 4.4 > 3$$

Now, i.e. the z-score is highly significant. Therefore, we reject $H_0$ i.e. it is not likely that the two samples are from the same population.

Example

Random samples from two populations gave the following results:

<table>
<thead>
<tr>
<th>Population A</th>
<th>Population B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>490</td>
</tr>
<tr>
<td>S.D.</td>
<td>50</td>
</tr>
<tr>
<td>Size</td>
<td>300</td>
</tr>
</tbody>
</table>

Is the difference between the means significant?

Solution: $H_o : \mu_1 = \mu_2$, the difference between the two means is not significant. i.e. $H_a : \mu_1 > \mu_2$.

We have $\bar{x}_1 = 490, \bar{x}_2 = 500, \sigma_1 = 50, \sigma_2 = 40, n_1 = 300$ and $n_2 = 300$.

$$|z| = \left| \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right| = \left| \frac{490 - 500}{\sqrt{\frac{50^2}{300} + \frac{40^2}{300}}} \right| = \left| \frac{-10}{\sqrt{\frac{2500}{300} + \frac{1600}{300}}} \right|$$

$$= \left| \frac{-10}{\sqrt{\frac{4100}{300}}} \right| = \left| \frac{-10}{\sqrt{13.66}} \right| = \left| \frac{-10}{3.70} \right| = 2.7315 < 1.96$$

Therefore, the null hypothesis $H_0$ is accepted i.e. the difference between the two means is not significant.
8.15 Test for difference between Proportions

If two samples are drawn from different populations, we may be interested in finding out whether the difference between the proportion of successes is significant or not. Let \( x_1 \) and \( x_2 \) be the number of items possessing the attribute A, in the random sampling of sizes \( n_1 \) and \( n_2 \) from two populations respectively.

Then the sample proportions of successes are \( \hat{P}_1 = \frac{x_1}{n_1} \) and \( \hat{P}_2 = \frac{x_2}{n_2} \), if \( P_1 \) and \( P_2 \) are proportion of successes in the two populations and \( Q_1 = 1 - P_1, Q_2 = 1 - P_2 \) then

\[
\varepsilon = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{P_1 Q_1}{n_1} + \frac{P_2 Q_2}{n_2}}}
\]

Under the hypothesis that the proportions in two populations are equal.

i.e. \( P_1 = P_2 = P \) \( Q_1 = Q_2 = Q \) (say) then

\[
\varepsilon = \frac{(\hat{P}_1 - \hat{P}_2) - (P_1 - P_2)}{\sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}}} = \frac{\hat{P}_1 - \hat{P}_2}{\sqrt{\frac{PQ}{n_1} + \frac{PQ}{n_2}}}
\]

In general, however, we do not know the population’s proportion of success. In such a case we can replace \( P \) by its best estimate \( \hat{P} = \) the pooled estimate of the actual proportion in the population, where

\[
\text{Pooled estimate (} P) = \frac{n_1 \hat{P}_1 + n_2 \hat{P}_2}{n_1 + n_2}
\]

\[
\hat{P} = \frac{\hat{P}_1 + \hat{P}_2}{n_1 + n_2}
\]

Example A machine produced 20 defective articles in a batch of 500. After overhauling it produced 3 defective in a batch of 100. Has the machine improved?

Solution: \( H_0 : P_1 = P_2 \) i.e. The machine has not improved after overhauling \( H_a : P_1 \neq P_2 \)

Now \( P_1 = \frac{20}{500} = 0.032 \) and \( P_2 = \frac{3}{100} = 0.030 \)
Two Tailed and One Tailed Tests

While testing a hypothesis, we often talk of two-tailed tests and one-tailed tests. In the previous tests the critical region lay along both the tails of the distributions. That is, we did not want sample statistic (say mean) to be away from the population parameter (say mean) in either direction. The test for such a hypothesis is non-directional or two-sided or two-tailed. A two-tailed test of hypothesis will reject the null hypothesis $H_0$, if the sample statistic is significantly higher than or lower than the hypothesized population parameter. Thus in two-tailed test the rejection (critical) region is located in both the tails.

For example, suppose you suspect that a particular 6th grader’s performance on a test in Mathematics is not a true representative of the students who have appeared. The national mean score in this test was found to be 75. The alternative (or research) hypothesis is:

$$H_a : \mu \neq 75$$

while the null hypothesis is:

$$H_0 : \mu = 75.$$ 

Now our pre-determined probability level is 95% i.e. 5% level of significance for this test. Both tests have the rejection (or critical) region of 5% i.e. 0.05. Now this rejection region is divided between both the tails of the distribution (see figure 1) i.e. 2.5% or 0.25 in the upper tail and 2.5% or 0.25 in the upper tail and 2.5% or 0.25 in the lower tail since your hypothesis gives only a difference and not a direction. You will reject the null hypothesis on the basis that the sample mean falls into the area
beyond 1.96 S.E. Otherwise if it falls into area 0.475 corresponds to 1.96 S.E. you can accept the null hypothesis.

Suppose you want to reduce the risk of committing a Type I error, reduce the size of the rejection region. If the hypothesis is treated at 1% i.e. 0.01 level of significance. If we consult the table of areas under the normal curve (table - 2), we find the acceptance region of 0.495 (one half of .99) is equal to 2.58 S.E. from $H_0$ i.e. $z$-score = 0.

You will still reject the null hypothesis of no difference, if the class sample is either much higher or much lower than our population mean of 75. As distinguished from the two-tailed test, we can apply a directional - one sided i.e. one-tailed test also because in some cases it is necessary to guard against only small values of $\bar{x}$ (i.e. sample mean). One-tailed test is so called because the rejection region will be located in only one-tail, which may either be on the upper or the lower side of the distribution depending upon the alternative ($H_a$) hypothesis formula. For example, we want to test a hypothesis that the average income per household is greater than $\text{₦}5000$ against the alternative hypothesis that the income is $\text{₦}10000$ or more. We will place all $\square$ risk on the upper-side of the theoretical
sampling distribution and the test will be one-tailed. On the other hand, if we are testing that the average income per household is N5000 against $H_a$ that the income is less than N5000 or less, the risk is on the lower side of the distribution and the test will be one sided.

Summing up, if the population as specified mean say $\mu_0$ then the null hypothesis would be $H_0: \mu = \mu_0$ and alternative (researcher’s) hypothesis could be:

1. $H_a: \mu > \mu_0$ (i.e. $\mu > \mu_0$ or $\mu < \mu_0$)
2. $H_a: \mu > \mu_0$
3. $H_a: \mu < \mu_0$

**Test of Significance for Small Samples**

So far we have discussed problems belonging to large samples. When a small sample (size < 30) is considered, the above tests are inapplicable because the assumptions we made for large sample tests, do not hold good for small samples. In case of small samples it is not possible to assume:

i. that the random sampling distribution of a statistics normal and

ii. the sample values are sufficiently close to population values to calculate the S.E. of estimate.

Thus an entirely new approach is required to deal with problems of small samples. But one should note that the methods and theory of small samples are applicable to large samples but its converse is not true.

**Degree of freedom (df):** By degree of freedom we mean the number of classes to which the value can be assigned arbitrarily or at will without voicing the restrictions or limitations placed.
For example, we are asked to choose any 4 numbers whose total is 50. Clearly we are at freedom to choose any 3 numbers say 10, 23, 7 but the fourth number, 10 is fixed since the total is 50 \([50 - (10 + 23 + 7) = 10]\). Thus we are given a restriction; hence the freedom of selection of number is \(4 - 1 = 3\). The degree of freedom (df) is denoted by \(\nu\) (nu) or df and it is given by \(\nu = n - k\), where \(n\) = number of classes and \(k\) = number of independent constrains (or restrictions).

In general for a Binomial distribution, \(\nu = n - 1\)

For Poisson distribution, \(\nu = n - 2\) (since we use total frequency and arithmetic mean).

For normal distribution, \(\nu = n - 3\) (since we use total frequency, mean and standard deviation) etc.

This is the same as testing for \(\mu\) when \(\sigma^2\) is unknown and or the population is normal or \(n \leq 30\) (small sample). Here is the hypothesis setting:

\[
\begin{align*}
\text{Ho} & \quad \mu = \mu_0 & \text{Reject Ho, if} & \quad |t| > t_{\alpha/2(n-1)} \\
\mu_0 & \quad \mu > \mu_0 & \quad t > t_{\alpha(n-1)} \\
\mu_0 & \quad \mu < \mu_0 & \quad t < -t_{\alpha(n-1)} \\
\end{align*}
\]

\[
t = \frac{x - \mu}{s\sqrt{n}} \sim t(n-1)
\]

**Example 3**: In an intelligence test on ten students, the following scores were obtained: 105, 102, 90, 85, 130, 110, 120, 115, 125, 100. Given that the average score for the class `before` special tuition for the test was 105, has the special tuition improves the performance. Use \(\alpha = 1\%\).

**Solution**

\[
\begin{align*}
\text{Ho: } \mu &= 105 \\
\text{H1: } \mu &> 105 \\
\alpha &= 1\%
\end{align*}
\]

From the data, mean \(x = 108.2\), standard deviation \(s = 14.65\)

236
Therefore, test statistic is

\[ t = \frac{108.2 - 105}{14.65/\sqrt{10}} = 1.96 \]

Tabulated \( t = 2.82 \)

**Conclusion:** Since \( 1.96 < 2.82 \), accept Ho. There is no evidence to show that a special tuition have improved students’ performance.

**Post-Test**

1. Define the following with respect to hypothesis testing:
   i. composite hypothesis
   ii. alternate hypothesis
   iii. null hypothesis
   iv. type I error
   v. type II error

2. Explain a one-tailed and two-tailed test with the aid of appropriate diagrams.

3. A random sample of 200 commercial banks shows a mean profit of \( \text{N}120\text{m} \) with a standard deviation of \( \text{N}25\text{m} \). Find the probability that the mean profit of commercial banks is at least:
   a. \( \text{N}125\text{m} \)
   b. \( \text{N}110\text{m} \)
   Test at 1% significance level.

4. The manageress of Christie supermarket stated that the average number of snacks sold per day was 250 with a variance of 225. A worker in the supermarket wants to test the accuracy of the claims. The worker took samples of the sales for 25 days and the average number of snacks sold per day was 243. Using 0.05 level of significance, should the claim of the manageress be accepted?

5. Explain the following statistical terms as they relate to hypothesis testing.
   a. Critical region
   b. Type I error
   c. Level of significance
6. State clearly the null and alternative hypothesis of test of two means for one-tailed test and two-tailed test.

7. It is claimed that the Distance Learning graduates employed by Procter and Gamble commit average of 13 calculation errors per week. A random check conducted on ten of them resulted in the following average number of errors per week: 15, 19, 22, 17, 13, 16, 12, 24, 20, 14. Test the hypothesis at 95% confidence level that the claim is an understatement.

8. An Economist in 2001 compared the wages of Federal Government and State Government employees. He discovered that the mean monthly wage of Federal Government employees is N1,000 more than that of State Government employees. He repeated the study in 2006, since he wonders whether or not there is still the same difference in their pay. He finds that the mean monthly wage of 100 Federal Government employees, chosen at random, is N17,000 with a standard deviation of N2000, while a random sample of 81 state Government employees has a mean wage of N15,500 with a standard deviation of N1,000. At a 0.05 level of significant, should the economist conclude that the difference in wages of Federal Government and State Government employees is unchanged?

9. A stockbroker claims that he can predict with 80% accuracy whether a stock market value will rise or fall during the coming month. As a test, he predicts the outcome of 40 stocks and is correct in 28 of the predictions. Does this evidence support the stockbroker’s claim?

10. A sample survey conducted among the inhabitant of Ibadan recently shows that 627 of 800 persons interviewed prefer to live in the suburb areas of Ibadan. Test the hypothesis that the true proportion of persons who prefer to live in the suburb is 0.75.

11. In a big city, 325 men out of 600 men were found to be smokers. Does this information support the conclusion that the majority of men in this city are smokers? State the hypothesis clearly.
References

Arinola, J.A: Handout on Business Statistics written for Delta State University, Abraka, (Ibadan Study Centre).

ICAN ATS Statistics Handout

LECTURE THIRTEEN

The Chi – Square Distribution

Introduction
Generally speaking, the chi-square test is a statistical test used to examine differences with categorical variables. There are a number of features of the social world we characterize through categorical variables - religion, political preference, psychology, sociology, etc. To examine hypotheses using such variables, use the chi-square test.

Objectives
At the end of this lecture, you are expected to have mastered the following:
1. the meaning of chi-square distribution
2. characteristics of $\chi^2$-test;
3. assumptions of $\chi^2$-test;
4. uses of $\chi^2$ test;
5. test of goodness of fit;
6. test for independence of attributes; and
7. test for the population variance.

Pre-Test
1. When should the correction for continuity be used?
2. A die is suspected of being biased. It is rolled 24 times with the following result.

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>
3. Conduct a significance test to see if the die is biased.

4. When can you use either a Chi Square or a z test and reach the same conclusion?

5. Ten subjects are each given two flavors of ice-cream to taste and say whether they like them. One of the 10 liked the first flavor and seven out of 10 liked the second flavor. Can the Chi square test be used to test whether this difference in proportions (.10 versus .70) is significant? Why or why not?

6. A recent experiment investigated the relationship between smoking and urinary incontinence. Of 322 subjects in the study who were incontinent, 113 were smokers, 51 were former smokers, and 158 had never smoked. Of 284 control subjects who were not incontinent, 68 were smokers, 23 were former smokers, and 193 had never smoked. Do a significance test to see if there is a relationship between smoking and incontinence.

The Chi-square Test

The chi-square, denoted by the Greek letter, is frequently used in testing of hypothesis concerning the difference between a set of observed frequencies of a sample and a corresponding set of expected or theoretical frequencies. In other words, a test statistic which measures the discrepancy between observed or actual frequencies $O_1, O_2, \ldots, O_n$ and their corresponding expected frequencies $E_1, E_2, \ldots, E_n$ is called the chi-square ($\chi^2$) statistic. It is an important non-parametric distribution – free test. The advantage of the test based on the chi square distribution is that it can be generalized to more complex situations.

The chi-square test is used in two similar but distinct circumstances:

- for estimating how closely an observed distribution matches an expected distribution - we'll refer to this as the goodness-of-fit test
- for estimating whether two random variables are independent.

Before the explanation, let us examine the characteristics of $\chi^2$ - test.
Characteristics of $\chi^2$–test.
This distribution is similar to t- distribution where the critical values vary with the degree of freedom. For every increase in the number of the degrees of freedom there is a new $\chi^2$–distribution.

This possesses additive property so that when $\chi^2_1$ and $\chi^2_2$ are independent and have a chi-square distribution with $n_1$ and $n_2$ degrees of freedom $\chi^2_1 + \chi^2_2$ will also be distributed as a chi-square distribution with $n_1 + n_2$ degrees of freedom.

Where the degree of freedom is 30 and less the distribution of $\chi^2$ is skewed, But, for degrees of freedom greater than 30 in a distribution, the values of $\chi^2$ are normally distributed.

Chi-square distribution itself is a continuous function, but where the frequencies involved are small, the discreteness of small numbers introduce an error similar to that of the binomial as a continuous functions with small values of n. To overcome this difficulty, if any cell frequency is less than 10. Yate’s correction should be used.

The $\chi^2$ distribution has only one parameter, the number of degrees of freedom. The $\chi^2$ distribution is positively skewed to the right, especially when the number of degrees of freedom is small. As ‘v’ the degree of freedom increases, the distribution becomes less skewed and rapidly approaches symmetry. For a large ‘v’ the $\chi^2$ distribution is approximately normal, in any case it is continuous and unimodal.

Assumptions in $\chi^2$ test
The sample observation should be independent, that is, no individual item should be included twice or more in the sample.

The total number of observations should be reasonably large, say, more than 50.

No theoretical frequency should be small. Small is a relative term.. Preferably, each theoretical frequency should be in this range, $5 \leq$ frequency $\leq \infty$.

Uses of $\chi^2$–test
The important uses of $\chi^2$ test are:
As a test of goodness of fit,
As a test for independence of attributes,
As a test for a specified standard deviation.
The Goodness of Fit test

If one is given a set of observed frequencies obtained in an experiment and we want to test whether the experimental results support a particular hypothesis. A test for significance was developed by Karl Pearson in 1900. This test is known as \( \chi^2 \)-test of goodness of fit and is used to test if the deviations between observed and theoretical values can be attributed to chance (fluctuations of sampling) or due to some inadequacy of the theory to fit the observed data. Under the null hypothesis there is no significant difference between the observed and theoretical values, i.e., there is good compatibility between theory and experiment. Karl Pearson proved that the statistic

\[
\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}
\]

The key idea of the chi-square test is a comparison of observed and expected values. How many of something were expected and how many were observed in some process?

With these sets of figures, we calculate the chi-square statistic as follows:

\[
\chi^2 = \sum \frac{(\text{observed } \times \text{ frequency} - \text{expected } \times \text{ frequency})^2}{\text{expected } \times \text{ frequency}}
\]

Using this formula with the values in the table above gives us a value of 13.6.

Lastly, to determine the significance level we need to know the "degrees of freedom." In the case of the chi-square goodness-of-fit test, the number of degrees of freedom is equal to the number of terms used in calculating chi-square minus one. There were six terms in the chi-square for this problem - therefore, the number of degrees of freedom is five.

We then compare the value calculated in the formula above to a standard set of tables. The value returned from the table is 1.8%. We interpret this as meaning that if the die was fair (or not loaded), then the chance of getting a \( \chi^2 \) statistic as large or larger than the one calculated above is only 1.8%. In other words, there's only a very slim chance that these rolls came from a fair die. The Missouri Master is in serious trouble.
Recap
To recap the steps used in calculating a goodness-of-fit test with chi-square:

1. Establish hypotheses.
2. Calculate chi-square statistic. Doing so requires knowing:
   • The number of observations
   • Expected values
   • Observed values
3. Assess significance level. Doing so requires knowing the number of degrees of freedom.
4. Finally, decide whether to accept or reject the null hypothesis.

Testing Independence
The other primary use of the chi-square test is to examine whether two variables are independent or not. What does it mean to be independent, in this sense? It means that the two factors are not related. Typically in social science research, we're interested in finding factors that are related - education and income, occupation and prestige, age and voting behavior. In this case, the chi-square can be used to assess whether two variables are independent or not.

More generally, we say that variable Y is "not correlated with" or "independent of" the variable X if more of one is not associated with more of another. If two categorical variables are correlated their values tend to move together, either in the same direction or in the opposite.

Example
In this example, we want to know whether boys or girls get into trouble more often in school. Below is the table documenting the percentage of boys and girls who got into trouble in school:

<table>
<thead>
<tr>
<th>Got in Trouble</th>
<th>No Trouble</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>46</td>
<td>71</td>
</tr>
<tr>
<td>Girls</td>
<td>37</td>
<td>83</td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>154</td>
</tr>
</tbody>
</table>
To examine statistically whether boys got in trouble in school more often, we need to frame the question in terms of hypotheses.

1. **Establish Hypotheses**
   As in the goodness-of-fit chi-square test, the first step of the chi-square test for independence is to establish hypotheses. The null hypothesis is that the two variables are independent - or, in this particular case that the likelihood of getting in trouble is the same for boys and girls. The alternative hypothesis to be tested is that the likelihood of getting in trouble is not the same for boys and girls.

**Cautionary Note**
It is important to keep in mind that the chi-square test only tests whether two variables are independent. It cannot address questions of which is greater or less. Using the chi-square test, we cannot evaluate directly the hypothesis that boys get in trouble more than girls; rather, the test (strictly speaking) can only test whether the two variables are independent or not.

2. **Calculate the expected value for each cell of the table**
   As with the goodness-of-fit example described earlier, the key idea of the chi-square test for independence is a comparison of observed and expected values. How many were expected and how many were observed in some process? In the case of tabular data, however, we usually do not know what the distribution should look like (as we did with rolls of dice). Rather, in this use of the chi-square test, expected values are calculated based on the row and column totals from the table.

   The expected value for each cell of the table can be calculated using the following formula:

   \[
   \frac{\text{Row total} \times \text{Column total}}{\text{Total } \eta \text{ for table}}
   \]

   For example, in the table comparing the percentage of boys and girls in trouble, the expected count for the number of boys who got in trouble is:

   \[
   = \frac{(\text{Total number of boys} \times \text{Total number of students who got in trouble})}{(\text{Total } \eta \text{ for table})}
   \]
The first step, then, in calculating the chi-square statistic in a test for independence is generating the expected value for each cell of the table. Presented in the table below are the expected values (in parentheses and italics) for each cell:

<table>
<thead>
<tr>
<th></th>
<th>Got in Trouble</th>
<th>No Trouble</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>46 (40.97)</td>
<td>71 (76.02)</td>
<td>117</td>
</tr>
<tr>
<td>Girls</td>
<td>37 (42.03)</td>
<td>83 (77.97)</td>
<td>120</td>
</tr>
<tr>
<td>Total</td>
<td>83</td>
<td>154</td>
<td>237</td>
</tr>
</tbody>
</table>

3. **Calculate Chi-square statistic**  
With these sets of figures, we calculate the chi-square statistic as follows:

\[
\chi^2 = \sum \frac{(observed \times frequency - expected \times frequency)^2}{(expected \times frequency)}
\]

In the example above, we get a chi-square statistic equal to:

\[
\chi^2 = \frac{(46 - 40.97)^2}{40.97} + \frac{(37 - 42.03)^2}{42.03} + \frac{(71 - 76.03)^2}{76.03} + \frac{(83 - 77.97)^2}{77.97}
\]

\[
\chi^2 = 1.87
\]

4. **Assess significance level**  
Lastly, to determine the significance level we need to know the "degrees of freedom." In the case of the chi-square test of independence, the number of degrees of freedom is equal to the number of columns in the table minus one multiplied by the number of rows in the table minus one.

In this table, there were two rows and two columns. Therefore, the number of degrees of freedom is:

\[
\chi^2 = \frac{(46 - 40.97)^2}{40.97} + \frac{(37 - 42.03)^2}{42.03} + \frac{(71 - 76.03)^2}{76.03} + \frac{(83 - 77.97)^2}{77.97}
\]

\[
\chi^2 = 1.87
\]
We then compare the value calculated in the formula above to a standard set of tables. The value returned from the table is $p < 0.20$. Thus, we cannot reject the null hypothesis and conclude that boys are not significantly more likely to get in trouble in school than girls.

Recap
To recap the steps used in calculating a goodness-of-fit test with chi-square:

1. Establish hypotheses
2. Calculate expected values for each cell of the table.
3. Calculate chi-square statistic. Doing so requires knowing:
   a. The number of observations
   b. Observed values
4. Assess significance level. Doing so requires knowing the number of degrees of freedom

Finally, decide whether to accept or reject the null hypothesis

Chi-Square Test of Independence
The goodness-of-fit test discussed above is appropriate for situations that involve one categorical variable. If there are two categorical variables, and our interest is to examine whether these two variables are associated with each other, the chi-square ($\chi^2$) test of independence is the correct tool to use. This test is very popular in analyzing cross-tabulations in which an investigator is keen to find out whether the two attributes of interest have any relationship with each other.

The cross-tabulation is popularly called by the term “contingency table”. It contains frequency data that correspond to the categorical variables in the row and column. The marginal totals of the rows and columns are used to calculate the expected frequencies that will be part of the computation of the $\chi^2$ statistic. For calculations on expected frequencies, refer hyperstat on $\chi^2$ test.

If the columns are not contingent on the rows, then the rows and column frequencies are independent. The test of whether the columns are contingent on the rows is called the chi square test of independence. The null hypothesis is that there is no relationship between row and column frequencies.
Example: A marketing firm producing detergents is interested in studying the consumer behavior in the context of purchase decision of detergents in a specific market. This company is a major player in the detergent market that is characterized by intense competition. It would like to know in particular whether the income level of the consumers influence their choice of the brand. Currently there are four brands in the market. Brand 1 and Brand 2 are the premium brands while Brand 3 and Brand 4 are the economy brands.

A representative stratified random sampling procedure was adopted covering the entire market using income as the basis of selection. The categories that were used in classifying income level are: Lower, Middle, Upper Middle and High. A sample of 600 consumers participated in this study. The following data emerged from the study.

Cross Tabulation of Income versus Brand chosen (Figures in the cells represent number of consumers)

<table>
<thead>
<tr>
<th>Income</th>
<th>Brand1</th>
<th>Brand2</th>
<th>Brand3</th>
<th>Brand4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>25</td>
<td>15</td>
<td>55</td>
<td>65</td>
<td>160</td>
</tr>
<tr>
<td>Middle</td>
<td>30</td>
<td>25</td>
<td>35</td>
<td>30</td>
<td>120</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>50</td>
<td>55</td>
<td>20</td>
<td>22</td>
<td>147</td>
</tr>
<tr>
<td>Upper</td>
<td>60</td>
<td>80</td>
<td>15</td>
<td>18</td>
<td>173</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>175</td>
<td>125</td>
<td>135</td>
<td>600</td>
</tr>
</tbody>
</table>

Analyze the cross-tabulation data above using chi-square test of independence and draw your conclusions.

Solution:
Null Hypothesis: There is no association between the brand preference and income level (These two attributes are independent).
Alternative Hypothesis: There is association between brand preference and income level (These two attributes are dependent).
Let us take a level of significance of 5%.
In order to calculate the $\chi^2$ value, you need to work out the expected frequency in each cell in the contingency table. In our example, there are 4 rows and 4 columns amounting to 16 elements. There will be 16 expected frequencies. For calculating expected frequencies, please go through hyperstat. Relevant data tables are given below:

**Observed Frequencies (These are actual frequencies observed in the survey)**

<table>
<thead>
<tr>
<th>Income</th>
<th>Brand1</th>
<th>Brand2</th>
<th>Brand3</th>
<th>Brand4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>25</td>
<td>15</td>
<td>55</td>
<td>65</td>
<td>160</td>
</tr>
<tr>
<td>Middle</td>
<td>30</td>
<td>25</td>
<td>35</td>
<td>30</td>
<td>120</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>50</td>
<td>55</td>
<td>20</td>
<td>22</td>
<td>147</td>
</tr>
<tr>
<td>Upper</td>
<td>60</td>
<td>80</td>
<td>15</td>
<td>18</td>
<td>173</td>
</tr>
<tr>
<td>Total</td>
<td>165</td>
<td>175</td>
<td>125</td>
<td>135</td>
<td>600</td>
</tr>
</tbody>
</table>

**Expected Frequencies (These are calculated on the assumption of the null hypothesis being true: That is, income level and brand preference are independent)**

<table>
<thead>
<tr>
<th>Income</th>
<th>Brand1</th>
<th>Brand2</th>
<th>Brand3</th>
<th>Brand4</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower</td>
<td>44.000</td>
<td>46.667</td>
<td>33.333</td>
<td>36.000</td>
<td>160.000</td>
</tr>
<tr>
<td>Middle</td>
<td>33.000</td>
<td>35.000</td>
<td>25.000</td>
<td>27.000</td>
<td>120.000</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>40.425</td>
<td>42.875</td>
<td>30.625</td>
<td>33.075</td>
<td>147.000</td>
</tr>
<tr>
<td>Upper</td>
<td>47.575</td>
<td>50.458</td>
<td>36.042</td>
<td>38.925</td>
<td>173.000</td>
</tr>
<tr>
<td>Total</td>
<td>165.000</td>
<td>175.000</td>
<td>125.000</td>
<td>135.000</td>
<td>600.000</td>
</tr>
</tbody>
</table>
**Note:** The fractional expected frequencies are retained for the purpose of accuracy. Do not round them.

**Calculation:**
Compute

$$\chi^2 = \sum \left( \frac{(O-E)^2}{E} \right).$$

There are 16 observed frequencies (O) and 16 expected frequencies (E). As in the case of the goodness of fit, calculate this $\chi^2$ value. In our case, the computed $\chi^2 = 131.76$ as shown below: Each cell in the table below shows $(O-E)^2/E$.

<table>
<thead>
<tr>
<th></th>
<th>Brand1</th>
<th>Brand2</th>
<th>Brand3</th>
<th>Brand4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Income</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lower</td>
<td>8.20</td>
<td>21.49</td>
<td>14.08</td>
<td>23.36</td>
</tr>
<tr>
<td>Middle</td>
<td>0.27</td>
<td>2.86</td>
<td>4.00</td>
<td>0.33</td>
</tr>
<tr>
<td>Upper Middle</td>
<td>2.27</td>
<td>3.43</td>
<td>3.69</td>
<td>3.71</td>
</tr>
<tr>
<td>Upper</td>
<td>3.24</td>
<td>17.30</td>
<td>12.28</td>
<td>11.25</td>
</tr>
</tbody>
</table>

and there are 16 such cells. Adding all these 16 values, we get $\chi^2 = 131.76$.

The critical value of $\chi^2$ depends on the degrees of freedom. The degrees of freedom = (the number of rows-1) multiplied by (the number of columns-1) in any contingency table. In our case, there are 4 rows and 4 columns. So the degrees of freedom = (4-1). (4-1) = 9. At 5% level of significance, critical $\chi^2$ for 9 d.f = 16.92. Therefore reject the null hypothesis and accept the alternative hypothesis.

The inference is that brand preference is highly associated with income level. Thus, the choice of the brand depends on the income strata. Consumers in different income strata prefer different brands. Specifically, consumers in upper middle and upper income group prefer premium brands while consumers in lower income and middle-income category prefer economy brands. The company should develop suitable strategies to position its detergent products. In the marketplace, it should position...
economy brands to lower and middle-income category and premium brands to upper middle and upper income category.

The formula for chi square is:

$$\chi^2 = \sum \frac{(|E - O| - 0.5)^2}{E}$$

where $\chi^2$ is the symbol for the chi square, $E$ is an expected cell frequency, and $O$ is an observed cell frequency. The $\Sigma$ symbol is summation notation and means to sum up the quantity over both cells. Therefore, the formula says that for each cell you

1. take the absolute value of the difference between the expected cell frequency and the observed cell frequency
2. subtract 0.5 (the correction for continuity)
3. square the result,
4. divide by the expected frequency,
5. and finally, sum up the values across the cells.

For the present data,

$$\chi^2 = \frac{(|50 - 62| - 0.5)^2}{50} + \frac{(|50 - 38| - 0.5)^2}{50} = \frac{11.5^2}{50} + \frac{11.5^2}{50} = 5.29.$$  

The degree of freedom for chi square is equal to the number of categories minus one. For this section in which there are always just two categories (success and failure for the present example), the degrees of freedom is always one. A chi square table can be used to find that the two-tailed probability value for a chi square of 5.29 with one degree of freedom is 0.0214.

At the beginning of this section it was stated that the chi square test for proportions was equivalent to the one based on the normal distribution. It turns out that chi square will always equal $z^2$. For the present example, the value of $z$ was 2.3 and the value of chi square was 5.29. Note that $2.3^2 = 5.29$. The probability values for a $z$ of 2.3 and a chi square of 5.29 are identical ($p = 0.0214$).

Contingency tables are used to examine the relationship between subjects' scores on two qualitative or categorical variables. For example,
consider the hypothetical experiment on the effectiveness of early childhood intervention programs described in another section. In the experimental group, 73 of 85 students graduated from high school. In the control group, only 43 of 82 students graduated. These data are depicted in the contingency table shown below.

<table>
<thead>
<tr>
<th></th>
<th>Graduated</th>
<th>Failed to Graduate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental</strong></td>
<td>73</td>
<td>12</td>
<td>85</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>43</td>
<td>39</td>
<td>82</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>116</td>
<td>51</td>
<td>167</td>
</tr>
</tbody>
</table>

The cell entries are cell frequencies. The top left cell with a "73" in it means that 73 subjects in the experimental condition went on to graduate from high school: 12 subjects in the experimental condition did not. The table shows that subjects in the experimental condition were more likely to graduate than were subjects in the control condition. Thus, the column a subject is in (graduated or failed to graduate) is contingent upon (depends on) the row the subject is in (experimental or control condition). Therefore, 116/167 graduated. If the null hypothesis were true, the expected frequency for the first cell would equal the product of the number of people in the experimental condition (85) and the proportion of people graduating (116/167). This is equal to $(85)(116)/167 = 59.042$. Therefore, the expected frequency for this cell is 59.042. The general formula for expected cell frequencies is:

$$E_{ij} = \frac{T_i \times T_j}{N}$$

where $E_{ij}$ is the expected frequency for the cell in the $i$th row and the $j$th column, $T_i$ is the total number of subjects in the $i$th row, $T_j$ is the total number of subjects in the $j$th column, and $N$ is the total number of subjects in the whole table.
The calculations are shown below.

\[
\begin{align*}
E_{11} &= \frac{85 \times 116}{167} \\
E_{12} &= \frac{85 \times 51}{167} \\
E_{21} &= \frac{82 \times 116}{167} \\
E_{22} &= \frac{82 \times 51}{167}
\end{align*}
\]

Once the expected cell frequencies are computed, it is convenient to enter them into the original table as shown below. The expected frequencies are in parentheses.

<table>
<thead>
<tr>
<th></th>
<th>Graduated</th>
<th>Failed to Graduate</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Experimental</strong></td>
<td>73 (59.042)</td>
<td>12 (25.958)</td>
<td>85</td>
</tr>
<tr>
<td><strong>Control</strong></td>
<td>43 (56.958)</td>
<td>39 (25.042)</td>
<td>82</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>116</td>
<td>51</td>
<td>167</td>
</tr>
</tbody>
</table>

The formula for chi square test for independence is

\[
\chi^2 = \sum \frac{(O - E)^2}{E}
\]

For this example,

\[
\begin{align*}
E_{11} &= \frac{85 \times 116}{167} \\
E_{12} &= \frac{85 \times 51}{167} \\
E_{21} &= \frac{82 \times 116}{167} \\
E_{22} &= \frac{82 \times 51}{167}
\end{align*}
\]

\[
\chi^2 = 22.01.
\]

The degrees of freedom are equal to (R-1) (C-1) where R is the number of rows and C is the number of columns. In this example, R = 2 and C = 2, so df = (2-1) (2-1) = 1. A chi square table can be used to determine that for df = 1, a chi square of 22.01 has a probability value less than 0.0001.
In a table with two rows and two columns, the chi square test of independence is equivalent to a test of the difference between two sample proportions. In this example, the question is whether the proportion graduating from high school differs as a function of condition. Whenever the degrees of freedom equal one (as they do when \( R = 2 \) and \( C = 2 \)), chi square is equal to \( z^2 \). Note that the test of the difference between proportions for these data results in a \( z \) of 4.69 which, when squared, equals 22.01.

The same procedures are used for analyses with more than two rows and/or more than two columns. For example, consider the following hypothetical experiment: A drug that decreases anxiety is given to one group of subjects before they attempted to play a game of chess against a computer. The control group was given a placebo. The contingency table is shown below.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Win</th>
<th>Lose</th>
<th>Draw</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>12 (14.29)</td>
<td>18 (14.29)</td>
<td>10 (11.43)</td>
<td>40</td>
</tr>
<tr>
<td>Placebo</td>
<td>13 (10.71)</td>
<td>7 (10.71)</td>
<td>10 (8.57)</td>
<td>30</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>25</td>
<td>20</td>
<td>70</td>
</tr>
</tbody>
</table>

The expected frequencies are shown in parentheses. As in the previous example, each expected frequency is computed by multiplying the row total by the column total and dividing by the total number of subjects. For example, the expected frequency for the "Drug-Lose" condition is the product of the row total (40) and the column total (25) divided by the total number of subjects (70): \( (40)(25)/70 = 14.29 \).

The chi square is calculated using the formula:

\[
\chi^2 = \sum \frac{(E - O)^2}{E}
\]

\[
\chi^2 = \frac{(14.29 - 12)^2}{14.29} + \frac{(14.29 - 18)^2}{14.29} + \ldots + \frac{(8.57 - 10)^2}{8.57} = 3.52.
\]
The df are \((R-1) (C-1) = (2-1)(3-1) = 2\). A chi square table shows that the probability of a chi square of 3.52 with 2 degrees of freedom is 0.172. Therefore, the effect of the drug is not significant.

**Summary of Computations**

1. Create a table of cell frequencies.
2. Compute row and column totals.
3. Compute expected cell frequencies using the formula:
   \[ E_{ij} = \frac{T_i \times T_j}{N} \]
   where \(E_{ij}\) is the expected frequency for the cell in the ith row and the jth column, \(T_i\) is the total number of subjects in the ith row, \(T_j\) is the total number of subjects in the jth column, and \(N\) is the total number of subjects in the whole table.
4. Compute Chi Square using the formula:
   \[ \chi^2 = \sum \frac{(E - O)^2}{E} \]
5. Compute the degrees of freedom using the formula: \(df = (R-1) (C-1)\) where \(R\) is the number of rows and \(C\) is the number of columns.
6. Use a chi square table to look up the probability value.

Note that the correction for continuity is not used in the chi square test of independence.

**Assumptions of the Chi Square Test of Independence**

A key assumption of the chi square test of independence is that each subject contributes data to only one cell. Therefore the sum of all cell frequencies in the table must be the same as the number of subjects in the experiment.

Consider an experiment in which each of 12 subjects threw a dart at a target once using his or her preferred hand and once using his or her non-preferred hand. The data are shown below:

<table>
<thead>
<tr>
<th></th>
<th>Hit</th>
<th>Missed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred hand</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>Non-preferred hand</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>
It would not be valid to use the chi square test of independence on these data since each subject contributed data to two cells: one cell based on their performance with their preferred hand and one cell based on their performance with their non-preferred hand. The total of the cell frequencies in the table is 24 but the total number of subjects is only 12.

**Summary**
Already you’ve learnt about the operations of chi-square test and its applications. Consult other textbooks for post test on this topic.
LECTURE FOURTEEN

Time series Analysis

Introduction
Time series is a sequence of numerical data points in successive order, usually occurring in uniform intervals. In plain English, a time series is simply a sequence of numbers collected at regular intervals over a period of time. Time series analysis can be useful to see how a given asset, security, unemployment or economic variable changes overtime or how it changes compared to other variables over the same time period.

Objectives
At the end of this lecture, it is expected that the following would have been mastered successfully:
1. meaning and components of time series;
2. methods of analyzing time series data;
3. estimation of trends;
4. moving average techniques;
5. method of least square regression; and
6. seasonal variation and forecasting.

Pre-Test
1. What is time series?
2. Name the four components of time series (i.e., the four patterns exhibited by time series).
3. What is the appropriate type of forecasting method to use in each of the following scenarios:
i. Standard brands have developed a new type of outdoor paint. The company wishes to forecast sales based on new housing starts.

ii. Holiday Inn. Inc., is attempting to predict the demand next year for motel rooms, based on a history of demand observations.

iii. National income since 1960 to date.


4. Enumerate the goals of time series analysis.

5. Discuss the role of forecasting for the following functions of the firm:
   i. Marketing
   ii. Accounting
   iii. Production line
   iv. Finance
   v. Manpower planning

6. The expenditure (Nm, or millions of Naira) on buildings and equipment in a state of the country is given in the table below:

<table>
<thead>
<tr>
<th>Year</th>
<th>Expenditure</th>
</tr>
</thead>
<tbody>
<tr>
<td>1994</td>
<td>200</td>
</tr>
<tr>
<td>1995</td>
<td>203</td>
</tr>
<tr>
<td>1996</td>
<td>207</td>
</tr>
<tr>
<td>1997</td>
<td>220</td>
</tr>
<tr>
<td>1998</td>
<td>242</td>
</tr>
<tr>
<td>1999</td>
<td>256</td>
</tr>
<tr>
<td>2000</td>
<td>257</td>
</tr>
<tr>
<td>2001</td>
<td>273</td>
</tr>
<tr>
<td>2002</td>
<td>305</td>
</tr>
<tr>
<td>2003</td>
<td>341</td>
</tr>
</tbody>
</table>
What is a Time Series?

A time series is a collection of observations of well-defined data items obtained through repeated measurements over time. For example, measuring the value of retail sales each month of the year would comprise a time series. This is because sales revenue is well defined, and consistently measured at equally spaced intervals. Data collected irregularly or only once are not time series.

Time series methods are often called naïve methods, as they require no information other than the past values of the variable being predicted. The term time series is just a fancy term for a collection of observations of some economic or physical phenomenon drawn at discrete points in time, usually spaced. The idea is that information can be inferred from the pattern of past observations and can be used to forecast future values of the series. Examples of time series are the annual GDP, daily sales of pure water. In fact, most data in macroeconomics and finance come in the form of time series – a set of repeated observations of the same variable, such as Gross National Product (GNP) or a stock return. Economics - weekly share prices, monthly profits. Sociology - crime figures (number of arrests, etc), employment figures. Meteorology - daily rainfall, wind speed, temperature.

In statistics, signal processing, and many other fields, a time series is a sequence of data points, measured typically at successive times, spaced at (often uniform) time intervals. Time series analysis comprises methods that attempt to understand such time series, often either to understand the

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>379</td>
</tr>
<tr>
<td>2005</td>
<td>395</td>
</tr>
<tr>
<td>2006</td>
<td>426</td>
</tr>
<tr>
<td>2007</td>
<td>449</td>
</tr>
<tr>
<td>2008</td>
<td>452</td>
</tr>
</tbody>
</table>

i. Draw a line graph of the data.

ii. Calculate a three-year moving average and plot it on your graph.

iii. Explain why moving averages are not satisfactory for predicting the trend of expenditure during the period 2009 – 2014.
underlying context of the data points (where did they come from? what generated them?), or to make forecasts (predictions). **Time series forecasting** is the use of a model to forecast future events based on known past events: to forecast future data points before they are measured. A standard example in econometrics is the opening price of a share of stock based on its past performance.

The term time series analysis is used to distinguish a problem, firstly from more ordinary data analysis problems (where there is no natural ordering of the context of individual observations), and secondly from spatial data analysis where there is a context that observations (often) relate to geographical locations. There are additional possibilities in the form of space-time models (often called spatial-temporal analysis). A time series model will generally reflect the fact that observations close together in time will be more closely related than observations further apart. In addition, time series models will often make use of the natural one-way ordering of time so that values in a series for a given time will be expressed as deriving in some way from past values, rather than from future values (see time reversibility.)

Methods for time series analyses are often divided into two classes: frequency-domain methods and time-domain methods. The former centre around spectral analysis and, recently, wavelet analysis, and can be regarded as model-free analyses well-suited to exploratory investigations. Time-domain methods have a model-free subset consisting of the examination of auto-correlation and cross-correlation analysis, but it is here that partly and fully-specified time series models make their appearance.

Two goals of time series analysis are:

i. Identifying patterns represented by a sequence of observations.

ii. Forecasting future values.

A time series is a sequence of observations which are ordered in time (or space). If observations are made on some phenomenon throughout time, it is most sensible to display the data in the order in which they arose, particularly since successive observations will probably be dependent. Time series are best displayed in a scatter plot. The series value \( X \) is plotted on the vertical axis and time \( t \) on the horizontal axis.
Time is called the independent variable (in this case however, something over which you have little control). There are two kinds of time series data:

1. Continuous, where we have an observation at every instant of time, e.g. lie detectors, electrocardiograms. We denote this using observation X at time t, X(t).
2. Discrete, where we have an observation at (usually regularly) spaced intervals. We denote this as Xt.

**Components of Time Series**
We want to increase our understanding of a time series by picking out its main features. One of these main features is the trend component. Descriptive techniques may be extended to forecast (predict) future values.

**Trend** is a long term movement in a time series. It is the underlying direction (an upward or downward tendency) and rate of change in a time series, when allowance has been made for the other components.

A simple way of detecting trend in seasonal data is to take averages over a certain period. If these averages change with time we can say that there is evidence of a trend in the series. There are also more formal tests to enable detection of trend in time series.

It can be helpful to model trend using straight lines, polynomials etc.

**Cyclical Component**
Another feature of a time series is the cyclical component. A series is cyclical if it appears to oscillate about the trend curve for a long time. The movement is assumed to be periodical, e.g. the appearance of a comet. It is described as the periodic rise and fall of the whole series over a number of years.

In weekly or monthly data, the cyclical component describes any regular fluctuations.

It is a non-seasonal component which varies in a recognizable cycle.
Seasonal Component

The seasonal component or movement is another feature of time series. A series is said to be seasonal if it appears to follow identical pattern for the same period for a long time. The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude. It arises from systematic, calendar related influences such as Natural conditions, business and administrative procedures, e.g. start and end of school term/semester, Social and cultural behaviour. Examples are: the sales of ice cream during dry season, sales of umbrella during rainy season and, sales of greeting cards during yuletide periods. It also includes calendar related systematic effects that are not stable in their annual timing or are caused by variations in the calendar from year to year, such as trading day effects and moving holiday effects e.g. Easter and Ileya and Idel moulud.

In weekly or monthly data, the seasonal component, often referred to as seasonality, is the component of variation in a time series which is dependent on the time of year. It describes any regular fluctuations with a period of less than one year. For example, the costs of various types of fruits and vegetables, unemployment figures and average daily rainfall, all show marked seasonal variation.

We are interested in comparing the seasonal effects within the years, from year to year; removing seasonal effects so that the time series is easier to cope with; and, also interested in adjusting a series for seasonal effects using various models.

Irregular Component

One of these main features is the irregular component (or 'noise'). This is an erratic movement of a series which is usually caused by occurrence of unusual events such as war, earthquake or workers strike, fires, flood etc.

The irregular component is that left over when the other components of the series (trend, seasonal and cyclical) have been accounted for.

Time series decomposition seeks to separate each of these elements, quantify its value and prepare a forecast combining all the elements. Of the various elements, the most important are the trend and seasonal variation.
Analysis of time series Data

Time series data are analyzed for various reasons, such as estimation of the trend to smooth a highly seasonal series and/or for the purpose of forecasting.

Smoothing

Smoothing techniques are used to reduce irregularities (random fluctuations) in time series data. They provide a clearer view of the true underlying behaviour of the series.

In some time series, seasonal variation is so strong it obscures any trends or cycles which are very important for the understanding of the process being observed. Smoothing can remove seasonality and makes long term fluctuations in the series stand out more clearly.

The most common type of smoothing technique is moving average smoothing although others do exist. Since the type of seasonality will vary from series to series, so must the type of smoothing.

Exponential Smoothing

Exponential smoothing is a smoothing technique used to reduce irregularities (random fluctuations) in time series data, thus providing a clearer view of the true underlying behaviour of the series. It also provides an effective means of predicting future values of the time series (forecasting).

Moving Average Smoothing

A moving average is a form of average which has been adjusted to allow for seasonal or cyclical components of a time series. Moving average smoothing is a smoothing technique used to make the long term trends of a time series clearer.

When a variable, like the number of unemployed, or the cost of strawberries, is graphed against time, there are likely to be considerable seasonal or cyclical components in the variation. These may make it difficult to see the underlying trend. These components can be eliminated by taking a suitable moving average.
By reducing random fluctuations, moving average smoothing makes long term trends clearer.

**Running Medians Smoothing**
Running medians smoothing is a smoothing technique analogous to that used for moving averages. The purpose of the technique is the same, to make a trend clearer by reducing the effects of other fluctuations.

**Differencing**
Differencing is a popular and effective method of removing trend from a time series. This provides a clearer view of the true underlying behaviour of the series.

**Autocorrelation**
Autocorrelation is the correlation (relationship) between members of a time series of observations, such as weekly share prices or interest rates, and the same values at a fixed time interval later. More technically, autocorrelation occurs when residual error terms from observations of the same variable at different times are correlated (related)

**Extrapolation**
Extrapolation is when the value of a variable is estimated at times which have not yet been observed. This estimate may be reasonably reliable for short times into the future, but for longer times, the estimate is liable to become less accurate.

**What are Seasonal Effects?**
A seasonal effect is a systematic and calendar related effect. Some examples include the sharp escalation in most Retail series which occurs around December in response to the Christmas period, or an increase in water consumption in summer due to warmer weather. Other seasonal effects include trading day effects (the number of working or trading days in a given month differs from year to year which will impact upon the level of activity in that month) and moving holidays (the timing of
holidays such as Easter varies, so the effects of the holiday will be experienced in different periods each year).

What is Seasonal Adjustment and Why Do We Need It?
Seasonal adjustment is the process of estimating and then removing from a time series influences that are systematic and calendar related. Observed data needs to be seasonally adjusted as seasonal effects can conceal both the true underlying movement in the series, as well as certain non-seasonal characteristics which may be of interest to analysts.

Why can't we just compare Original Data from the same period in each year?
A comparison of original data from the same period in each year does not completely remove all seasonal effects. Certain holidays such as Easter and Chinese New Year fall in different periods in each year, hence they will distort observations. Also, year to year values will be biased by any changes in seasonal patterns that occur over time. For example, consider a comparison between two consecutive March months i.e. compare the level of the original series observed in March for 2000 and 2001. This comparison ignores the moving holiday effect of Easter. Easter occurs in April for most years but if Easter falls in March, the level of activity can vary greatly for that month for some series. This distorts the original estimates. A comparison of these two months will not reflect the underlying pattern of the data. The comparison also ignores trading day effects. If the two consecutive months of March have different composition of trading days, it might reflect different levels of activity in original terms even though the underlying level of activity is unchanged. In a similar way, any changes to seasonal patterns might also be ignored. The original estimates also contain the influence of the irregular component. If the magnitude of the irregular component of a series is strong compared with the magnitude of the trend component, the underlying direction of the series can be distorted.

However, the major disadvantage of comparing year to year original data is lack of precision and time delays in the identification of turning points in a series. Turning points occur when the direction of underlying level of the series changes, for example when a consistently decreasing
series begins to rise steadily. If we compare year apart data in the original series, we may miss turning points occurring during the year. For example, if March 2001 has a higher original estimate than March 2000, by comparing these year apart values, we might conclude that the level of activity has increased during the year. However, the series might have increased up to September 2000 and then started to decrease steadily.

When is Seasonal Adjustment Inappropriate?

When a time series is dominated by the trend or irregular components, it is nearly impossible to identify and remove what little seasonality is present. Hence seasonally adjusting a non-seasonal series is impractical and will often introduce an artificial seasonal element.

How Do we Identify Seasonality?

Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend. The following diagram depicts a strongly seasonal series. There is an obvious large seasonal increase in December retail sales in New South Wales due to Christmas shopping. In this example, the magnitude of the seasonal component increases over time, as does the trend.

Time Series Models and Graphs

Decomposition models are typically additive or multiplicative, but can also take other forms such as pseudo-additive. The standard graph for a time series is a line graph/diagram, known technically as a historigram. It is obtained by plotting the time series values (on the vertical axis) against time (on the horizontal axis) 0 as a single point which are joined by straight line segments.

Additive Decomposition

In some time series, the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls. In such cases, an additive model is appropriate.
In the additive model, the observed time series \( (O_t) \) is considered to be the sum of three independent components: the seasonal \( S_t \), the trend \( T_t \) and the irregular \( I_t \).

\[ \text{Observed series} = \text{Trend} + \text{Seasonal} + \text{Irregular} \]

That is

\[ O_t = T_t + S_t + I_t \]

Each of the three components has the same units as the original series. The seasonally adjusted series is obtained by estimating and removing the seasonal effects from the original time series. The estimated seasonal component is denoted by \( \hat{S}_t \). The seasonally adjusted estimates can be expressed by:

\[ \text{Seasonally adjusted series} = \text{Observed series} - \text{Seasonal} = \text{Trend} + \text{Irregular} \]

In symbols,

\[ SA_t = O_t - \hat{S}_t = T_t + I_t \]

The following figure depicts a typically additive series. The underlying level of the series fluctuates but the magnitude of the seasonal spikes remains approximately stable.
**Multiplicative Decomposition**

In many time series, the amplitude of both the seasonal and irregular variations increase as the level of the trend rises. In this situation, a multiplicative model is usually appropriate.

In the multiplicative model, the original time series is expressed as the product of trend, seasonal and irregular components.

\[ \text{Observed series} = \text{Trend} \times \text{Seasonal} \times \text{Irregular} \]

or

\[ O_t = T_t \times S_t \times I_t \]

The seasonally adjusted data then becomes:

\[ \text{Seasonally Adjusted series} = \frac{\text{Observed}}{\text{Seasonal}} = \text{Trend} \times \text{Irregular} \]
or

\[ SA_t = \frac{D_t}{T_t \times I_t} \]

Under this model, the trend has the same units as the original series, but the seasonal and irregular components are unit-less factors, distributed around 1.

Most of the series analysed by the ABS show characteristics of a multiplicative model. As the underlying level of the series changes, the magnitude of the seasonal fluctuations varies as well.

**Figure 5: Monthly NSW ANZ Job Advertisements**

![Graph showing monthly NSW ANZ Job Advertisements]

**Pseudo-Additive Decomposition**

The multiplicative model cannot be used when the original time series contains very small or zero values. This is because it is not possible to divide a number by zero. In these cases, a pseudo additive model combining the elements of both the additive and multiplicative models is
used. This model assumes that seasonal and irregular variations are both dependent on the level of the trend but independent of each other.

The original data can be expressed in the following form:

\[ O_t = T_t \times S_t \times (1 - 1) + T_t \times (1 - 1) \]

\[ = T_t \times (S_t + I_t - 1) \]

The pseudo-additive model continues the convention of the multiplicative model to have both the seasonal factor \( S_t \) and the irregular factor \( I_t \) centred around one. Therefore we need to subtract one from \( S_t \) and \( I_t \) to ensure that the terms \( T_t \times (S_t - 1) \) and \( T_t \times (I_t - 1) \) are centred on zero. These terms can be interpreted as the additive seasonal and additive irregular components respectively and because they are centred around zero the original data \( O_t \) will be centred around the trend values \( T_t \).

The seasonally adjusted estimate is defined to be:

\[ \hat{S}A_t = O_t \times \hat{T}_t \times (\hat{S}_t \times 1) \]

\[ = \hat{T}_t \times \hat{I}_t \]

where \( \hat{T}_t \) and \( \hat{S}_t \) are the trend and seasonal component estimates. In the pseudo-additive model, the trend has the same units as the original series, but the seasonal and irregular components are unit-less factors, distributed around 1.

An example of series that requires a pseudo-additive decomposition model is shown below. This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.
How do I know which Decomposition Model to use?

To choose an appropriate decomposition model, the time series analyst will examine a graph of the original series and try a range of models, selecting the one which yields the most stable seasonal component. If the magnitude of the seasonal component is relatively constant regardless of changes in the trend, an additive model is suitable. If it varies with changes in the trend, a multiplicative model is the most likely candidate. However if the series contains values close or equal to zero, and the magnitude of seasonal component appears to be dependent upon the trend level, then pseudo-additive model is most appropriate.

Methods of Calculating Trend

The trend can be estimated by:

a. Semi – averages
b. Method of least squares (regression method)
c. Moving average technique
d. Free hand method.
a. Semi – Averages

Procedure
i. Split the data into a lower and upper group
ii. Find the mean value of each group
iii. Plot on a graph, each mean against an appropriate time point.
iv. The line joining the two plotted points is the required trend

b. Method of least squares (regression method)

Procedure
i. Take the physical time points as values (coded as 1, 2, 3, etc if necessary) of the independent variable.
ii. Take the data values themselves as values of the dependent variable y.
iii. Calculate the least squares regression line of y on x, y = a + bx.
iv. Translate the regression line as t = a + bx, where any given value of time point x will yield at corresponding value of the trend, t.

c. Moving Average technique

Moving averages (of period n) for the values of a time series are arithmetic means of successive and overlapping values, taken n at a time.

Example: 1
Daily takings (₦’000) of a supermarket over three successive weeks are shown in the table.

Supermarket takings (₦’000)

<table>
<thead>
<tr>
<th>Day</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>92</td>
<td>89</td>
<td>87</td>
</tr>
<tr>
<td>Tuesday</td>
<td>98</td>
<td>93</td>
<td>90</td>
</tr>
<tr>
<td>Wednesday</td>
<td>104</td>
<td>100</td>
<td>97</td>
</tr>
<tr>
<td>Thursday</td>
<td>106</td>
<td>103</td>
<td>99</td>
</tr>
</tbody>
</table>
a. Draw a time chart of the data.

b. Use an appropriate moving average to find the trend of the series and plot it on your time chart. Comment on the trend.

**Solution**

<table>
<thead>
<tr>
<th>Day</th>
<th>Takings ('000)</th>
<th>7 – day total</th>
<th>7 – day moving average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mon</td>
<td>92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tue</td>
<td>98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wed</td>
<td>104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thu</td>
<td>106</td>
<td>753</td>
<td>107.57</td>
</tr>
<tr>
<td>Fri</td>
<td>120</td>
<td>750</td>
<td>107.14</td>
</tr>
<tr>
<td>Sat</td>
<td>132</td>
<td>745</td>
<td>106.43</td>
</tr>
<tr>
<td>Sun</td>
<td>101</td>
<td>741</td>
<td>105.86</td>
</tr>
<tr>
<td>Mon</td>
<td>89</td>
<td>738</td>
<td>105.43</td>
</tr>
<tr>
<td>Tue</td>
<td>93</td>
<td>733</td>
<td>104.71</td>
</tr>
<tr>
<td>Wed</td>
<td>100</td>
<td>726</td>
<td>103.71</td>
</tr>
<tr>
<td>Thu</td>
<td>103</td>
<td>724</td>
<td>103.43</td>
</tr>
<tr>
<td>Fri</td>
<td>115</td>
<td>722</td>
<td>103.14</td>
</tr>
<tr>
<td>Sat</td>
<td>125</td>
<td>719</td>
<td>102.71</td>
</tr>
<tr>
<td>Sun</td>
<td>99</td>
<td>716</td>
<td>102.29</td>
</tr>
<tr>
<td>Mon</td>
<td>87</td>
<td>712</td>
<td>101.71</td>
</tr>
<tr>
<td>Tue</td>
<td>90</td>
<td>711</td>
<td>101.57</td>
</tr>
<tr>
<td>Wed</td>
<td>97</td>
<td>707</td>
<td>101.00</td>
</tr>
<tr>
<td>Thu</td>
<td>99</td>
<td>705</td>
<td>100.71</td>
</tr>
<tr>
<td>Fri</td>
<td>114</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sat</td>
<td>121</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sun</td>
<td>97</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The trend is slowly downwards, almost linear.

Example 2:
The numbers of passengers traveling through a regional airport have been recorded quarterly as shown in the table.

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter 1</th>
<th>Quarter 2</th>
<th>Quarter 3</th>
<th>Quarter 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>12</td>
<td>15</td>
<td>21</td>
<td>13</td>
</tr>
<tr>
<td>2001</td>
<td>15</td>
<td>19</td>
<td>27</td>
<td>17</td>
</tr>
<tr>
<td>2002</td>
<td>21</td>
<td>28</td>
<td>36</td>
<td>16</td>
</tr>
<tr>
<td>2003</td>
<td>22</td>
<td>32</td>
<td>44</td>
<td>25</td>
</tr>
<tr>
<td>2004</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimate the trend of passenger numbers by calculating centred four-quarterly moving averages. Comment on any irregularity in the trend.

Use multiplicative model to estimate average seasonal indices for each quarter. Describe, in non-technical language, the fluctuation between quarters.
### Solution

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>Passengers (x100,000)</th>
<th>4-Quarter Moving Average</th>
<th>Trend</th>
<th>Actual/Trend (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>1</td>
<td>12</td>
<td>15.25</td>
<td>15.625</td>
<td>134.40</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>15</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>21</td>
<td>16.00</td>
<td>16.500</td>
<td>78.79</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>13</td>
<td>17.00</td>
<td>17.750</td>
<td>84.51</td>
</tr>
<tr>
<td>2001</td>
<td>1</td>
<td>15</td>
<td>18.50</td>
<td>19.000</td>
<td>100.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>19</td>
<td>19.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>27</td>
<td>21.00</td>
<td>20.250</td>
<td>133.33</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>17</td>
<td>23.25</td>
<td>22.125</td>
<td>76.84</td>
</tr>
<tr>
<td>2002</td>
<td>1</td>
<td>21</td>
<td>25.50</td>
<td>24.375</td>
<td>86.15</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>28</td>
<td>25.25</td>
<td>25.375</td>
<td>110.34</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>36</td>
<td>25.50</td>
<td>26.000</td>
<td>141.87</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>16</td>
<td>26.50</td>
<td>27.500</td>
<td>61.54</td>
</tr>
<tr>
<td>2003</td>
<td>1</td>
<td>22</td>
<td>28.50</td>
<td>29.625</td>
<td>80.00</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>32</td>
<td>30.75</td>
<td>31.125</td>
<td>108.02</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>44</td>
<td>31.50</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2004</td>
<td>1</td>
<td>25</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The trend rises until the second quarter of 2002, and then it rises again after a short period of being constant.

The average seasonal variations (%) are calculated as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>Q1</th>
<th>Q2</th>
<th>Q3</th>
<th>Q4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>134.40</td>
<td>78.79</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001</td>
<td>84.51</td>
<td>100.00</td>
<td>133.33</td>
<td>76.84</td>
</tr>
<tr>
<td>2002</td>
<td>86.15</td>
<td>110.34</td>
<td>141.87</td>
<td>61.54</td>
</tr>
<tr>
<td>2003</td>
<td>80.00</td>
<td>108.02</td>
<td>141.37</td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>83.55</td>
<td>106.12</td>
<td>137.74</td>
<td>72.39</td>
</tr>
<tr>
<td>X</td>
<td>83.59</td>
<td>106.17</td>
<td>137.81</td>
<td>72.43</td>
</tr>
<tr>
<td>400/399.8</td>
<td>Sum = 400</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

275
In quarters 1 and 4, actual numbers are below trend, namely about 84% and 72% of trend respectively (i.e. 16% and 28% below). In quarters 2 and 3, actual numbers are above trend, namely about 106% and 138% of trend respectively (i.e. 6% and 38% above).

Example 2:
The quarterly revenues in Naira generated in 4 years by Nigmarship Soap making factory are as shown below:

<table>
<thead>
<tr>
<th>Quarters</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td>18</td>
<td>15</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>22</td>
<td>20</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>25</td>
<td>20</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>27</td>
<td>22</td>
<td>35</td>
<td></td>
</tr>
</tbody>
</table>

a. Compute the trend, using the four quarters centred moving average method. Correct your answers to 1 decimal place.

b. Use the ratio-to-trend method to compute the seasonal indices (unadjusted). Correct your answers to 2 decimal places.
Solution

<table>
<thead>
<tr>
<th>Y</th>
<th>4Q Moving Totals</th>
<th>2-4Q Moving Totals</th>
<th>4Q Centred Moving Average</th>
<th>Y/T</th>
</tr>
</thead>
<tbody>
<tr>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>99</td>
<td>100</td>
<td>25.0</td>
<td>0.60</td>
</tr>
<tr>
<td>28</td>
<td>105</td>
<td>106</td>
<td>25.8</td>
<td>1.09</td>
</tr>
<tr>
<td>40</td>
<td>110</td>
<td>115</td>
<td>26.9</td>
<td>1.49</td>
</tr>
<tr>
<td>22</td>
<td>114</td>
<td>124</td>
<td>28.0</td>
<td>0.79</td>
</tr>
<tr>
<td>20</td>
<td>119</td>
<td>133</td>
<td>29.1</td>
<td>0.69</td>
</tr>
<tr>
<td>32</td>
<td>122</td>
<td>141</td>
<td>30.1</td>
<td>1.06</td>
</tr>
<tr>
<td>45</td>
<td>122</td>
<td>144</td>
<td>30.5</td>
<td>1.48</td>
</tr>
<tr>
<td>25</td>
<td>122</td>
<td>144</td>
<td>30.5</td>
<td>0.82</td>
</tr>
<tr>
<td>20</td>
<td>125</td>
<td>147</td>
<td>30.9</td>
<td>0.65</td>
</tr>
<tr>
<td>32</td>
<td>127</td>
<td>152</td>
<td>31.5</td>
<td>1.02</td>
</tr>
<tr>
<td>48</td>
<td>129</td>
<td>156</td>
<td>32.0</td>
<td>1.50</td>
</tr>
<tr>
<td>27</td>
<td>132</td>
<td>161</td>
<td>32.6</td>
<td>0.83</td>
</tr>
<tr>
<td>22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Quarters

<table>
<thead>
<tr>
<th>Quarter Totals</th>
<th>Quarters means (unadjusted)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 2 3 4</td>
<td>1.49 0.79 0.69 1.06</td>
</tr>
<tr>
<td></td>
<td>1.48 0.82 0.65 1.02</td>
</tr>
<tr>
<td></td>
<td>1.50 0.83 - -</td>
</tr>
<tr>
<td></td>
<td>4.47 2.44 1.94 3.17</td>
</tr>
<tr>
<td></td>
<td>1.49 0.81 0.65 1.06</td>
</tr>
</tbody>
</table>
Post-Test

1. The table provides data about the use of electricity in a large college in Nigeria for each quarter from 1996 Qtr 1 to 2000 Qtr 2. The trend has been calculated by the method of moving averages. Calculate the average seasonal component for each type of quarter using an additive model.

<table>
<thead>
<tr>
<th>Year/Quarter</th>
<th>Units used ('000)</th>
<th>Trend ('000s of unit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1996/1</td>
<td>207</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>152</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>127</td>
<td>161.125</td>
</tr>
<tr>
<td>4</td>
<td>176</td>
<td>161.000</td>
</tr>
<tr>
<td>1997/1</td>
<td>172</td>
<td>166.000</td>
</tr>
<tr>
<td>2</td>
<td>186</td>
<td>166.875</td>
</tr>
<tr>
<td>3</td>
<td>133</td>
<td>171.000</td>
</tr>
<tr>
<td>4</td>
<td>177</td>
<td>173.625</td>
</tr>
<tr>
<td>1998/1</td>
<td>204</td>
<td>172.000</td>
</tr>
<tr>
<td>2</td>
<td>175</td>
<td>171.375</td>
</tr>
<tr>
<td>3</td>
<td>131</td>
<td>168.375</td>
</tr>
<tr>
<td>4</td>
<td>174</td>
<td>165.875</td>
</tr>
<tr>
<td>1999/1</td>
<td>183</td>
<td>166.125</td>
</tr>
<tr>
<td>2</td>
<td>176</td>
<td>166.375</td>
</tr>
<tr>
<td>3</td>
<td>132</td>
<td>167.750</td>
</tr>
<tr>
<td>4</td>
<td>175</td>
<td>167.750</td>
</tr>
<tr>
<td>2000/1</td>
<td>194</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>164</td>
<td></td>
</tr>
</tbody>
</table>

2. Explain what you understand by the following terms in relation to a time series:
   a. Trend;
   b. Seasonal component;
   c. Additive model;
   d. Multiplicative model
Explain when it would be appropriate to use an additive model rather than a multiplicative model for time series analysis.

3. The table shows the quarterly number of reported major road accidents within the boundaries of a large town from 2002 to 2004.

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>Accidents (y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2002</td>
<td>1</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>36</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>46</td>
</tr>
<tr>
<td>2003</td>
<td>1</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>47</td>
</tr>
<tr>
<td>2004</td>
<td>1</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>49</td>
</tr>
</tbody>
</table>

i. Draw a time chart of the data.

ii. Explain why it is appropriate to calculate a four-quarterly centred moving average to estimate the trend of accidents.

iii. Estimate the trend of accidents in this way and plot it on your chart.

iv. Comment on the trend.

v. Numbering the quarters from 2002 Quarter 1 to 2004 Quarter 4 in order from 1 to 12 and using these as x values, calculate the least squares regression line of Accidents (y) on Quarter number.

vi. Plot this line on your chart and compare it with the moving average trend.

4. a. What is a time series?
   b. List the four components of a time series.
   c. Which of these components are usually estimable?
   d. State the two time series models.
e. The monthly sales of Dudu Osun soap industry in a given year are as listed below:

<table>
<thead>
<tr>
<th>Month</th>
<th>Jan</th>
<th>Feb</th>
<th>March</th>
<th>April</th>
<th>May</th>
<th>June</th>
<th>July</th>
<th>Aug</th>
<th>Sept</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sales ('000)</td>
<td>90</td>
<td>92</td>
<td>94</td>
<td>110</td>
<td>112</td>
<td>104</td>
<td>107</td>
<td>111</td>
<td>121</td>
<td>123</td>
<td>124</td>
<td>12</td>
</tr>
</tbody>
</table>

i. Calculate a 3-month moving average for the sales to one decimal place.

ii. Plot the actual values and the trend values on the same graph.

5. a. Explain the term seasonal variation as used in time series analysis. Comment on any seasonal variation in question 4 above.

b. Identify the components of a business or economic time series. What is the main essence of time series analysis?

c. Values in million naira of personal cheques cleared between 2000 and 2005 are as follows:

<table>
<thead>
<tr>
<th>Year</th>
<th>1st</th>
<th>2nd</th>
<th>3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>127</td>
<td>114</td>
<td>134</td>
</tr>
<tr>
<td>2001</td>
<td>130</td>
<td>120</td>
<td>143</td>
</tr>
<tr>
<td>2002</td>
<td>133</td>
<td>123</td>
<td>149</td>
</tr>
<tr>
<td>2003</td>
<td>136</td>
<td>126</td>
<td>152</td>
</tr>
<tr>
<td>2004</td>
<td>133</td>
<td>120</td>
<td>152</td>
</tr>
<tr>
<td>2005</td>
<td>130</td>
<td>123</td>
<td>-</td>
</tr>
</tbody>
</table>

Compute the relevant seasonal indices for the series using percentage moving average method. Deseasonalize the data. Mention any other alternative method of computing the seasonal indices.
LECTURE FIFTEEN

Official Statistics

Introduction
Statistics arose, no later than the 18th century, from the need of states to collect data on their people and economies, in order to administer them. Its meaning broadened in the early 19th century to include the collection and analysis of data in general. Today statistics is widely employed in government, business, and the natural and social sciences.

Because of its origins in government and its data-centric world view, statistics is considered to be not a subfield of mathematics but rather a distinct field that uses mathematics.

Objectives
Students are expected to master the following after reading through this lecture:
1. meaning and aims of official statistics;
2. categories of official statistics;
3. common indicators used in official statistics;
4. users and producers of official statistics;
5. production and collection process of official statistics; and
6. official statistics presentation and quality criteria.

Pre-Test
1. What is meant by official statistics?
2. Enumerate the various categories of official statistics.
3. List out the main indicators of the following:
   i. Population
   ii. Gender
   iii. Employment
   iv. Energy
   v. Trade

4. What is economic statistics? What are the main indicators of economy?

5. List the types of official statistics

CONTENT

Meaning of official statistics

Official statistics are related directly to the field of statistics and cover all major areas of citizens' lives, such as economic and social development, living conditions, health, education, and the environment.

During the 16th and 17th centuries, statistics were a method for counting and listing populations and State resources. The term statistics comes from the Latin status (state) indicating that the origin of the profession related to the needs of governments. The term official statistics simply describes statistics from an organization or institution that officially belongs to the State. According to the Organization for Economic Cooperation and Development, OECD, "official statistics are statistics disseminated by the national statistical system, excepting those that are explicitly not to be official".

Another definition is given by Beimer and Lyberg in 2003
"Almost every country in the world has one or more government agencies (usually national institutes) that supply decision-makers and other users including the general public and the research community with a continuing flow of information (...). This bulk of data is usually called official statistics. Official statistics should be objective and easily accessible and produced on a continuing basis so that measurement of change is possible."

Official statistics result from the collection and processing of data into statistical information by the government institution responsible for
that subject-matter domain. They are then disseminated to stakeholders and the general public. Statistical information allows users to draw a relevant, reliable and accurate picture of the development of the country, compare differences between countries and changes over time. They enable stakeholders and decision makers to be well informed and develop policies for addressing actual development challenges.

*Official statistics make information on development accessible to the public and therefore assist in the accountability of public decision-making. One use of official statistics is to measure the impact of public policies and highlight the need for development.*

**Aim of Official Statistics**

Official statistics provide a picture of a country or different phenomena through data, and images such as graph and maps. Statistical information covers different subject areas (economic, demographic, social etc.). Statistical information serves a variety of uses for monitoring developments in a country as well as provides basic information for decision-making, evaluations and assessments at different levels, and - traditionally - governments. Official statistics are a key component of description and comparison of the situation and development of a country. They are also essential for non-material content such as unregistered migration and economic transactions.

The goal of national statistical agencies is to produce relevant, objective and accurate statistics so that they keep people and organizations well informed about the aim and its development. They answer the public and users demands for better access to information, assisting good policy and decision-making.

In addition, demand by users for more information has significantly increased since computing technologies have greatly helped in dealing with growing volumes of data. In recent years there has been an increasing demand of a broadening range of users to be able to access information sources quickly and easily. The Internet has been one answer to this request that is why statistical agencies have developed their abilities to satisfy more standardized delivery models to meet users' expectations.
Various Categories of Official Statistics

The Fundamental Principles of Official Statistics were adopted in 1992 by the United Nations Economic Commission for Europe, UNECE. According to the first Principle "the official statistical information is an essential basis for development in the economic, demographic, social and environment fields and for mutual knowledge and trade among the states and peoples of the region".

Following is a list of the different statistical categories and their subgroups:

**Demographic Statistics** concern the evolution of population and migration. They deal with the measurement of undeclared migration or measurement of emigration, migration flows, immigrant population, and foreign-born population.

**Social Statistics** provide data on society, on the population and all that surrounds it.

Demographic and social statistics include: population and migration, labour, education, health, income and consumption, social protection, human settlements and housing, justice and crime, culture, political and other community activities, time use, living conditions, poverty and cross-cutting social issues, information society.

**Gender and Special Population Groups Statistics** reflect the realities of the lives of women and men and policy issues related to gender such as domestic violence, violence against women, gender pay gap as well as elderly, disabled, minority groups.

**Economic Statistics** is the branch of statistics that studies the economic activities of a country through macroeconomics statistics, economic accounts, business statistics, globalization, sectoral statistics, agriculture-forestry-fisheries, energy, mining-manufacturing-construction, transport, tourism, banking-insurance-financial statistics, government finance-fiscal and public sector statistics, international trade and balance of payments, prices, labour cost, science and technology.

The category includes econometrics which is a combination of economics and statistics, whose aim is to analyze the economic relationship.
Environmental Statistics concerns the environmental field and its variability, especially environment and sustainable development.

Most common indicators used in official statistics are land use, water supply and consumption, environmental protection expenditure etc.

A country’s profile can be summarized using only figures and data classified in different categories. Statistical indicators provide an overview of the social, demographic and economic structure of the country. Moreover, these indicators enable making comparisons between countries on an international scale, with agreements on those indicators.

For population, the main indicators are:
- Total population
- Population density
- Population by age
- Life expectancy at birth and at age 65
- Foreign born
- Foreigners in population
- Total fertility rate
- Infant mortality

The gender statistics include:
- Women in labour force
- Gender pay gap

In the employment category:
- Employment rate
- Unemployment rate
- Youth unemployment rate
- Economic activity rate (women and men)
- Employment in major sectors: agriculture, industry, services
There are many indicators for the economy:

- Gross Domestic Product
- Gross Domestic Product per capita
- Real GDP growth rate
- GDP by major economic sectors: agriculture, industry, services
- Consumer price index
- Purchasing Power Parity (PPP)
- Exchange rate
- Gross external debt

For trade indicators we find:

- Exports of goods and services
- Imports of goods and services
- Balance of payments
- Trade balance
- Major import partners
- Major export partners

Environment indicators:

- Land use
- Water supply and consumption
- Environmental protection expenditure
- Generation and treatment of waste
- Chemical use

For the energy field:

- Total energy consumption
- Primary energy sources
- Energy consumption in transport
- Electricity consumption
- Consumption of renewable energy sources
The Three User Types of Official Statistics

Official statistics are intended for a wide range of users including governments (central and local), research institutions such as NISER, Development Policy Centre etc., professional statisticians, journalists and the media, businesses, educational institutions and the general public. There are three types of users: those with a general interest, business interest or research interest. Each of these user groups has different needs for statistical information.

Users with a General Interest

*Users with a general interest* include the media, schools and the general public. They use official statistics in order to be informed on a particular topic, to observe trends within the society of a local area, country, and region of the world.

Users with a Business Interest

*Users with a business interest* include decision makers and users with a particular interest for which they want more detailed information. For them, official statistics are an important reference, providing information on the phenomena or circumstances their own work is focusing on. For instance, those users will take some official statistics into consideration before launching a product, or deciding on a specific policy or on a marketing strategy. As with the general interest users, this group does not
usually have a good understanding of statistical methodologies, but they need more detailed information than the general users.

**Users with a Research Interest**

*Users with a research interest* are universities, consultants and government agencies. They generally understand something about statistical methodology and want to dig deeper into the facts and the statistical observations; they have an analytical purpose in inventing or explaining interrelations of causes and effects of different phenomena. In this field, official statistics are also used to assess a government’s policies.

One common point for all these users is their need to be able to trust the official information. They need to be confident that the results published are authoritative and unbiased. Producers of official statistics must maintain a reputation of professionalism and independence.

The statistical system must be free from interference that could influence decisions on the choice of sources, methods used for **data collection**, the selection of results to be released as official, and the timing and form of **dissemination**. Statistical business processes should be transparent and follow international standards of good practice.

Statistical programs are decided on an annual or multi-annual basis by governments in many countries. They also provide a way to judge the performance of the statistical system.

**Producers**

Official statistics are collected and produced by **national statistical institutes** (NSIs), in Nigeria, National Bureau of Statistics (NBS) is responsible for the collection of official statistics. They are responsible for producing and disseminating official statistical information, providing the highest quality data. The criteria for quality are: relevance and completeness, timeliness, accuracy, accessibility and clarity, cost efficiency, transparency, comparability and coherence, which are called **quality principles**. Their role is also to foster statistical literacy among important user groups and the general public.

The core tasks of NBS, for both centralized and decentralized systems, are investing user needs and filtering these for relevance. Then they transform the relevant user needs into measurable concepts to
facilitate data collection and dissemination. This is part of the **production process**.

Statistical agencies also determine the resources necessary for the various activities and ensure relevance by keeping permanent networks with various representatives of different types of users. Moreover statistical producers have to anticipate user needs when producing new types of statistics in order to speed up the process. It can be called the "antenna function". Another way to assess the quality of their services is to undertake periodically user satisfaction surveys.

Nevertheless, NBS are usually not the only official and national institutes to produce official statistics. Central banks, National Population Commission and some ministries or other central authorities may have statistical functions as well. Together, all producers of official statistics form the **National statistical system** of a country. In some countries, especially those that have a federal structure, producers of official statistics exist also at regional or even municipal levels.

The NBS is in charge of the coordination between statistical producers and of ensuring the coherence and compliance of the statistical system to the principles. In countries with less than 25 million inhabitants, the NBS should also have the exclusive responsibility for all household surveys and all business surveys for official statistics. This should improve efficiency and ensure confidentiality.

The NBS has to pay particular attention to ensuring that the information materials, the terminology and the metadata of the statistical results disseminated are coherent and understandable for non-users especially in the case of diverging results compiled from different sources. The NBS also has to decide which ones are the official statistics. Moreover, it ensures the dissemination platforms for all official statistical registers from which contact information for sample surveys is extracted. Finally, the NBS should support and advise other producers of official statistics and organize meetings with all of them.

The NBS has a major responsibility as its President/Director General represents the entire system of official statistics, both at the national and at international levels. Statistical producers perform advocacy work on official statistics and can provide advice and services, such as training activities in order to broaden the know-how accumulated in the agency.
Production Process

The usual production process of official statistics includes 6 steps:

1. programming phase
2. design phase
3. data collection
4. processing phase
5. dissemination
6. evaluation

1. The programming phase starts the process with investigations into the information needs of users (topic of the information needed, its period, accuracy and timeliness). A lot of information is gathered, and then filtered in order to focus on one specific activity. Official statistics can generate results that fulfill a great number of user needs and not only target a unique user group. Filtering allows selecting relevant information. These needs are translated into the best way of collecting data from respondents.

2. The design phase is when tests and statistical surveys are designed or redesigned and tested. The first surveys are called pilots. Tools and resources are also prepared to conduct the surveys and to implement them fully. This phase includes the definition of the results to be published as official.

3. Data collection through statistical surveys can be done through different processes: by mail, face-to-face interviews, telephone interviews, internet, sample survey, sampling frame.

4. The processing phase includes data entry, control, coding and editing. This phase is highly IT-dependent; CATI, Computer-assisted telephone interviewing and CAPI, Computer-assisted personal interviewing techniques are really useful in terms of speed. The results and the quality parameters have to be analysed carefully before passing to the next phase. This monitoring can be systematically done by tallying the same phenomenon with other sources at the aggregate level.

5. Dissemination is more than the mere release of the results and statistical products such as press releases, electronic dissemination on the Internet or hard-copy publication, to customers. It may
include publications related to the topic with more details or analytical content, or may target specific user groups. It includes the generation of additional results for specific user requests such as statistical services. This means that consequently, microdata or the final sets of data have to be stored and well documented for a considerable period.

6. An evaluation of the whole process is necessary to identify and make improvements in efficiency and quality of the process launched by the statistical producer; this helps for the next programming exercise.

Sometimes statistical producers can make ad hoc surveys but this is not recommended in the case of official statistics. The methodological information should be used in most of the surveys to obtain relevant results.

Another data life-cycle object model for statistical information systems was developed by the Statistical Office of the European Communities, Eurostat in 2003.

A proposal for a new generic statistical business process model was submitted in 2007 by Statistics New Zealand and the United Nations Economic Commission for Europe Secretariat to a joint UNECE/Eurostat/OECD work session on statistical metadata and is already used by several official statistical agencies even if it is still being developed. The aim of the generic model is to provide a better basis for the production process, to make it more complete and detailed than the previous model.

Collection process
There are three main ways to collect data, by surveys, registers and censuses.

Statistical survey or sample survey
A statistical survey or a sample survey is an investigation about the characteristics of a phenomenon by means of collecting data from a sample of the population and estimating their characteristics through the systematic use of statistical methodology.
The main advantage is the *direct control* over data content and the possibility to ask *oriented questions*. Another advantage is the *rapidity of process and publication* that is possible with computing techniques like CAPI or CATI.

A disadvantage is the *high cost* and the *variable quality of data* collected when non-response can cause biases or when respondents are not able or willing to give correct answers.

There are various survey methods that can be used such as direct interviewing, telephone, mail, online surveys. The respondents of surveys can be called *primary respondents* that correspond to individual, households and companies.

**Register**

A register is a database that is updated continuously for a specific purpose and from which statistics can be collected and produced. It contains information on a complete group of units.

An advantage is the *total coverage* even if collecting and processing represent *low cost*. It allows producing more detailed statistics than using surveys. Different registers can be combined and linked together on the basis of defined keys (personal identification codes, business identification codes, address codes etc.). Moreover, individual administrative registers are usually of high quality and very detailed.

A disadvantage is the *possible under-coverage* that can be the case if the incentive or the cultural tradition of registering events and changes are weak, if the classification principles of the register are not clearly defined or if the classifications do not correspond to the needs of statistical production to be derived from them.

There are different types of registers:

→ *Administrative registers* or *records* can help the NSI in collecting data. Using the existing administrative data for statistical production may be approved by the public because it can be seen as a cost efficient method; individuals and enterprises are less harassed by a response burden; data security is better as fewer people handle it and data have an electronic format.
→ **Private registers** such as registers operated by insurance companies and employer organizations can also be used in the production process of official statistics, providing there is an agreement or legislation on this.

→ **Statistical registers** are frequently based on combined data from different administrative registers or other data sources.

→ For businesses, it is often legally indispensable to be registered in their country to a **business register** which is a system that makes business information collection easier.

→ It is possible to find **agricultural registers** and **registers of dwellings**.

**Census**

Census is the complete enumeration of a population or groups at a point in time with respect to well-defined characteristics (population, production). Censuses are not subject to an updating process contrary to registers. The information has to be collected at a reference period. The census should be taken at regular intervals in order to have comparable information available. Therefore, most of the time a census is conducted every 5 or 10 years. The data is collected through questionnaires that are either mailed to respondents or completed by an enumerator visiting respondents. It can also be done today by the Internet or by automated telephone interviewing.

An advantage is for small areas or sub-units census may be the only information source on social, demographic or economic characteristics. Often, census results provide a basis for sampling frames used in forthcoming surveys.

The major disadvantages of censuses are the usually high costs of their planning and implementation. Also, different understandings or interpretations of the terminology used in census questionnaires can be a problem.

In 2005, the United Nations Economic and Social Council adopted a resolution urging: "Member States to carry out a population and housing census and to disseminate census results as an essential source of information for small area, national, regional and international planning and development; and to provide census results to national stakeholders as well as the United Nations and other appropriate intergovernmental
organizations to assist in studies on population, environment, and socio-economic development issues and programs”.

Even though different types of data collection exist, the best estimates are based on a combination of different sources providing the strengths and reducing the weakness of each individual source.

**Official Statistics Presentation**

Official statistics can be presented in different ways. Analytical texts and tables are the most traditional ways. **Graphs** and charts summarize data highlighting information content visually. They can be extremely effective in expressing key results, or illustrating a presentation. Sometimes a picture is worth a thousand words. Graphs and charts usually have a heading describing the topic.

There are different types of graphic but usually the data determine the type that is going to be used.

To illustrate changes over time, a **line graph** would be recommended. This is usually used to display variables whose values represent a regular progression.

Stacked bar chart showing the sectoral contribution to total business services growth, 2001-2005 for members of UNECE.
For categorical data, it is better to use a bar graph either vertical or horizontal. They are often used to represent percentages and rates and also to compare countries, groups or illustrate changes over time. The same variable can be plotted against itself for two groups. An example of this is the age pyramid.

**Pie chart** can be used to represent share of 100 per cent. Pie charts highlight the topic well only when there are few segments.

**Stacked bar charts**, whether vertical or horizontal, are used to compare compositions across categories. They can be used to compare percentage composition and are most effective for categories that add up to 100 per cent, which make a full stacked bar chart. Their use is usually restricted to a small number of categories.

**Tables** are a complement to related texts and support the analysis. They help to minimize numbers in the description and also eliminate the need to discuss small variables that are not essential. Tables rank data by order or other hierarchies to make the numbers easily understandable. They usually show the figures from the highest to the lowest.

Another type of visual presentation of statistical information is **thematic map**. They can be used to illustrate differences or similarities between geographical areas, regions or countries. The most common statistical map that is used is called the choropleth map where different shades of a colour are used to highlight contrasts between regions; darker colour means a greater statistical value. This type of map is best used for ratio data but for other data, proportional or graduated symbol maps, such as circles, are preferred. The size of the symbol increases in proportion to the value of the observed object.

**Release**

Official statistics are part of our everyday life. They are everywhere: in newspapers, on television and radio, in presentations and discussions. For most citizens, the media provide their only exposure to official statistics. **Television** is the primary news source for citizens in industrialized
countries, even if radio and newspapers still play an important role in the dissemination of statistical information. On the other hand newspapers and specialized economic and social magazines can provide more detailed coverage of statistical releases as the information on a specific theme can be quite extensive. Official statistics provides us with important information on the situation and the development trends in our society.

Users can gather information making use of the services of the National Statistical Offices. They can easily find it on the agency’s website. The development of computing technologies and the Internet has enabled users - businesses, educational institutions and households among others- to have access to statistical information. The Internet has become an important tool for statistical producers to disseminate their data and information. People are able to access information online. The supply of information from statistical agencies has increased. Today the advanced agencies provide the information on their websites in an understandable way, often categorized for different groups of users. Several glossaries have been set up by different organizations or statistical offices to provide more information and definitions in the field of statistics and consequently official statistics.

Quality Criteria to be Respected
The quality criteria of a national statistical office are the following: relevance, impartiality, dissemination, independence, transparency, confidentiality, international standards. There principles apply not only to the NSO but to all producers of official statistics. Therefore, not every figure reported by a public body should be considered as official statistics, but those produced and disseminated according to the principles. Adherence to these principles will enhance the credibility of the NSO and other official statistical producers and build public trust in the reliability of the information and results that are produced.

Relevance
Relevance is the first and most important principles to be respected for national statistical offices. When releasing information, data and official statistics should be relevant in order to fulfill the needs of users as well as both public and private sector decision makers. Production of official statistics is relevant if it corresponds to different user needs like public,
governments, businesses, research community, educational institutions, NGOs and international organizations or if it satisfies basic information in each area and citizen’s right to information.

**Impartiality**

Once the survey has been made, the NBS checks the quality of the results and then they have to be disseminated no matter what impact they can have on some users, whether good or bad. All should accept the results released by the NBS as authoritative. Users need to perceive the results as unbiased representation of relevant aspects of the society. Moreover, the impartiality principle implies the fact that NBS have to use understandable terminology for statistics’ dissemination, questionnaires and material published so that everyone can have access to their information.

**Dissemination**

In order to maximize dissemination, statistics should be presented in a way that facilitates proper interpretation and meaningful comparisons. To reach the general public and non-expert users when disseminating, NBS have to add explanatory comments to explain the significance of the results released and make analytical comments when necessary. There is a need to identify clearly what the preliminary, final and revised results are, in order to avoid confusion for users. All results of official statistics have to be publicly accessible. There are no results that should be characterized as official and for the exclusive use of the government. Moreover they should be disseminated simultaneously.

**Independence**

Users can be consulted by NBS but the decisions should be made by statistical bodies. Information and activities of producers of official statistics should be independent of political control. Moreover, NBS have to be free of any political interference that could influence their work and thus, the results. They should not make any political advice or policy-perspective comments on the results released at anytime, even at press conferences or in interviews with the media.
Transparency
The need for transparency is essential for NBS to gain the trust of the public. They have to expose to the public the methods they use to produce official statistics, and be accountable for all the decisions they take and the results they publish. Also, statistical producers should warn users of certain interpretations and false conclusions even if they try to be as precise as possible. Furthermore, the quality of the accurate and timely results must be assessed prior to release. But if errors in the results occur before or after the data revision, they should be directly corrected and information should be disseminated to the users at the earliest possible time. Producers of official statistics have to set analytical systems in order to change or improve their activities and methods.

Confidentiality
All data collected by the national statistical office must protect the privacy of individual respondents, whether persons or businesses. But on the contrary, government units such as institutions cannot invoke statistical confidentiality. All respondents have to be informed about the purpose and legal basis of the survey and especially about the confidentiality measures. The statistical office should not release any information that could identify an individual or group without prior consent. After data collection, replies should go back directly to the statistical producer, without involving any intermediary. Data processing implies that filled-in paper and electronic form with full names should be destroyed.

International standards
The use of international standards at the national level aims to improve international comparability for national users and facilitate decision-making, especially when controversial. Moreover, the overall structure, including concepts and definitions, should follow internationally accepted standards, guidelines or good practices. International recommendations and standards for statistical methods approved by many countries provide them with a common basis like the two standards of the International Monetary Fund, IMF, SDDS for Special Data Dissemination Standards and GDDS for General Data Dissemination System. Their aim is to guide countries in the dissemination of their economic and financial data.
to the public. One approved, these standards have to be observed by all producers of official statistics and not only by the NBS.

Summary
In this lecture, we have learnt some useful things on official statistics. We have seen that official statistics has process and it is been coordinated by National Bureau of Statistics (NBS). Some other producers of official statistics were mentioned. All these bodies collect data, process it and disseminate same to the users for the general populace to benefit therein.

Post-Test
1. What are the production processes of official statistics?
2. Write short note on:
   i. Programming phase
   ii. Design phase
   iii. Data collection phase
3. What is sample survey? What are the advantages and disadvantages of sample survey.
4. List out the various survey methods that can be used to collect the official statistics.
5. What is meant by register? List out the types of registers. What are the advantages and disadvantage of register method of collecting official statistics?
6. How do we present official statistics?
7. Enumerate the quality criteria of national statistical office
References


OECD. Online Glossary of Statistical Terms http://stats.oecd.org/glossary/index.htm

Biemer, Paul and Lyberg Lars (2003), "Introduction to Survey Quality" – Business & Economics – Wiley - Hardback


International Institute for Sustainable Development http://www.iisd.org/sd/


