Mount Kenya University

COURSE TITLE: INTRODUCTION SOCIAL STATISTICS

COURSE CODE: DED1204

LESSON 1

INTRODUCTION

INTRODUCTION

Purpose

To introduce the student to the world of statistics and to acquaint them with the role of statistics in Business.

Objectives

1) Define statistics and explain its uses.
2) Define Business Statistics
3) State limitations of statistics.
4) Explain why statistics is distrusted.
5) Distinguish between descriptive and inferential statistics.
6) Explain the types of variables.
7) State the levels and scales of measurement

1.1 What is Statistics?

The word 'statistics' is defined by Croxton and Cowden as follows:-

"The collection, presentation, analysis and interpretation of the numerical data."

This definition clearly points out four stages in a statistical investigation, namely:

1) Collection of data 2) Presentation of data 3) Analysis of data 4) Interpretation of data

In addition to this, one more stage i.e. organization of data is suggested

Definition:

Social statistics is the science of good decision making in the face of uncertainty and is used in many disciplines such as sociology, psychology financial analysis, econometrics, auditing, production and operations including services improvement, and marketing research..

1.2 Uses of Statistics

1. To present the data in a concise and definite form: Statistics helps in classifying and tabulating raw data for processing and further tabulation for end users.
2. To make it easy to understand complex and large data: This is done by presenting the data in the form of tables, graphs, diagrams etc., or by condensing the data with the help of means, dispersion etc.
3. For comparison: Tables, measures of means and dispersion can help in comparing different sets of data..
4. In forming policies: It helps in forming policies like a production schedule, based on the relevant sales figures. It is used in forecasting future demands.
5. Enlarging individual experiences: Complex problems can be well understood by statistics, as the conclusions drawn by an individual are more definite and precise than mere statements on facts.
6. In measuring the magnitude of a phenomenon:- Statistics has made it possible to count the population of a country, the industrial growth, the agricultural growth, the educational level (of course in numbers)

1.3 Limitations of Statistics

1. Statistics does not deal with individual measurements. Since statistics deals with aggregates of facts, it cannot be used to study the changes that have taken place in individual cases. For example, the wages earned by a single industry worker at any time, taken by itself is not a statistical datum. But the wages of workers of that industry can be used statistically. (2) class marks

2. Statistics cannot be used to study qualitative phenomenon like morality, intelligence, beauty etc. as these cannot be quantified.
3. Statistical results are true only on an average:- The conclusions obtained statistically are not universal truths. They are true only under certain conditions. This is because statistics as a science is less exact as compared to the natural science.
4. Statistical data, being approximations, are mathematically incorrect. Therefore, they can be used only if mathematical accuracy is not needed.
5. Statistics, being dependent on figures, can be manipulated and therefore can be used only when the authenticity of the figures has been proved beyond doubt..

1.3 Distrust of Statistics

A Paris banker said, "Statistics is like a miniskirt, it covers up essentials but gives you the ideas."

The term distrust of statistics mean lack of confidence in statistical statements and methods.

The following reasons make statistics vulnerable to manipulations.

1. Figures are convincing and, therefore people easily believe them.
2. They can be manipulated in such a manner as to establish foregone conclusions.
3. The wrong representation of even correct figures can mislead a reader. For example, John earned Ksh 400,000 in 1990 - 1991 and Jane earned Ksh 500,000. Reading this one would form the opinion that Jane is decidedly a better worker than John. However if we carefully examine the statement, we might reach a different conclusion as Jane's earning period is unknown to us. Thus while working with statistics one should not only avoid outright falsehoods but be alert to detect possible distortion of the truth.

1.4 Types of Statistics

Broadly speaking, statistics may be divided into two categories, ie descriptive and inferential statistics.

In most research conducted on groups of people, you will use both descriptive and inferential statistics to analyze your results and draw conclusions.

1.4.1 Descriptive Statistics

Descriptive statistics is the term given to the analysis of data that helps describe, show or summarize data in a meaningful way such that, for example, patterns might emerge from the data. Descriptive statistics do not, however, allow us to make conclusions beyond the data we have analyzed or reach conclusions regarding any hypotheses we might have made. They are simply a way to describe our data.

Descriptive statistics allow us to present data in a more meaningful way which allows simpler interpretation of the data. For example, if we had the results of 100 pieces of students' coursework, we may be interested in the overall performance of those students. We would also be interested in the distribution or spread of the marks. Descriptive statistics allow us to do this. There are two general types of statistic that are used to describe data:

- Measures of central tendency: these are ways of describing the central position of a frequency distribution for a group of data.
- A frequency distribution is a table used to describe a data set. It lists intervals or ranges of data values called data classes together with the number of data values from the set that are in each class.
- The three common measures of central tendency are the :
- Mean
- Median
- mode
- Measures of spread or variation: these are ways of summarizing a group of data by describing how spread out the scores are.
- Spread or variation in data set is the amount of difference between data values.

- The common measures of spread are:
    - Range
    - Quartiles
    - Absolute deviation
    - variance
    - standard deviation.
- When we use descriptive statistics it is useful to summarize our group of data using a combination of tabulated description (i.e. tables), graphical description (i.e. graphs and charts) and statistical commentary (i.e. a discussion of the results).

1.4.2 Inferential Statistics

Inferential statistics aim to make inferences from data in order to make conclusions that go beyond the data.

inferential statistics are used to make inferences about a population from a sample in order to make assumptions about the wider population and/or make predictions about the future.

For example, a Board of Examiners may want to compare the performance of 1000 students that completed an examination. Of these, 500 students are girls and 500 students are boys. The 1000 students represent our "population". Whilst we are interested in the performance of all 1000 students, girls and boys, it may be impractical to examine the marks of all of these students because of the time and cost required to collate all of their marks. Instead, we can choose to examine a "sample" of these students and then use the results to make generalizations about the performance of all 1000 students. For the purpose of our example, we may choose a sample size of 200 students. Since we are looking to compare boys and girls, we may randomly select 100 girls and 100 boys in our sample. We could then use this, for example, to see if there are

any statistically significant differences in the mean mark between boys and girls, even though we have not measured all 1000 students.

1.5 Common Mistakes Committed In Interpretation of Statistics

1. Bias: - Bias means prejudice or preference of the investigator, which creeps in consciously and unconsciously in proving a particular point.
2. Generalization:- Some times on the basis of little data available one could jump to a conclusion, which leads to erroneous results.
3. Wrong conclusion:- The characteristics of a group if attached to an individual member of that group, may lead us to draw absurd conclusions.
4. Incomplete classification:- If we fail to give a complete classification, the influence of various factors may not be properly understood.
5. There may be a wrong use of percentages.
6. Technical mistakes may also occur.
7. An inconsistency in definition can even exist.
8. Wrong causal inferences may sometimes be drawn.

Core text

S.P Gupta (2004): Introduction to statistical methods 23rd-ed: vikas publishing house New Delhi

2. Futher reading

• Saleemi N.A (1997), Statistics Simplified Reprinted January 2011: Nairobi, Saleemi Publication limited.

• Saleemi N.A (1992), Quantitative Techniques: Nairobi, Saleemi Publication limited.

LESSON TWO

Basic statistical concepts

Population

Refers to the complete set of observations of a given characteristics of interest. (The universe)

Sample

Is a subset of a population. Its either representative or non representative

Individual or objects where the characteristic of interest is being observed or interviewed are referred to as elementary units

The whole list of elementary list is known as sampling list

Census

This is a study where all the elements in the sampling frame are included in the survey

If only a fraction of sampling frame is considered such a study is known as sample survey

Parameter

It's a summary measure of a population

Statistic

A summary measure of a sample

Statistics

The collection, presentation, analysis and interpretation of the numerical data. This definition clearly points out four stages in a statistical investigation, namely:

1) Collection of data 2) Presentation of data
3) Analysis of data 4) Interpretation of data

1.2 Uses of Statistics

- To present the data in a concise and definite form : Statistics helps in classifying and tabulating raw data for processing and further tabulation for end users.
- To make it easy to understand complex and large data : This is done by presenting the data in the form of tables, graphs,

diagrams etc., or by condensing the data with the help of means, dispersion etc.

- For comparison: Tables, measures of means and dispersion can help in comparing different sets of data..
- In forming policies: It helps in forming policies like a production schedule, based on the relevant sales figures. It is used in forecasting future demands.
- Enlarging individual experiences: Complex problems can be well understood by statistics, as the conclusions drawn by an individual are more definite and precise than mere statements on facts.
- In measuring the magnitude of a phenomenon:- Statistics has made it possible to count the population of a country, the industrial growth, the agricultural growth, the educational level (of course in numbers)

1.3 Limitations of Statistics

- Statistics does not deal with individual measurements. Since statistics deals with aggregates of facts, it cannot be used to study the changes that have taken place in individual cases. For example, the wages earned by a single industry worker at any time, taken by itself is not a statistical datum. But the wages of workers of that industry can be used statistically. (2) class marks
- Statistics cannot be used to study qualitative phenomenon like morality, intelligence, beauty etc. as these cannot be quantified.
- Statistical results are true only on an average:- The conclusions obtained statistically are not universal truths. They are true only under certain conditions. This is because statistics as a science is less exact as compared to the natural science.

- Statistical data, being approximations, are mathematically incorrect. Therefore, they can be used only if mathematical accuracy is not needed.
- Statistics, being dependent on figures, can be manipulated and therefore can be used only when the authenticity of the figures has been proved beyond doubt..

1.4 Common Mistakes Committed In Interpretation of Statistics

- Bias: - Bias means prejudice or preference of the investigator, which creeps in consciously and unconsciously in proving a particular point.
- Generalization:- Some times on the basis of little data available one could jump to a conclusion, which leads to erroneous results.
- Wrong conclusion:- The characteristics of a group if attached to an individual member of that group, may lead us to draw absurd conclusions.
- Incomplete classification:- If we fail to give a complete classification, the influence of various factors may not be properly understood.
- There may be a wrong use of percentages.
- Technical mistakes may also occur.
- An inconsistency in definition can even exist.
- Wrong causal inferences may sometimes be drawn.

1.6 variable

A variable is a characteristic of interest e.g object or individual that can be quantified

If immeasurable the character if interest is known as an attribute

1.6 Types of Variables

1.6.1 Discrete Variable

A *discrete variable* It is a variable that can only increase or decrease by full number not by fraction e.g 1, 2, 3, 4, and 5

1.6.2Continuous Variable

A variable that can theoretically assume an infinite number of values between any two points.

e.g Height

1.7 Scales of measurement

Measurement refers to assigning numbers to objects according a set of rules

1.7.1 Nominal Scale

Nominal measurement consists of assigning items to groups or categories. No quantitative information is conveyed and no ordering of the items is implied. Nominal scales are therefore qualitative rather than quantitative. Religious preference, race, and sex are all examples of nominal scales. Frequency distributions are usually used to analyze data measured on a nominal scale. The main statistic computed is the mode. Variables measured on a nominal scale are often referred to as categorical or qualitative variables.

1.7.2 Ordinal Scale

Measurements with ordinal scales are ordered in the sense that higher numbers represent higher values. However, the intervals between the numbers are not necessarily equal. For example, on a five-point rating scale measuring attitudes toward gun control, the difference between a rating of 2 and a rating of 3 may not represent the same difference as the difference between a rating of 4 and a rating of 5. There is no "true" zero point for ordinal scales since the zero point is chosen arbitrarily. The lowest point on the rating scale in the example was arbitrarily chosen to be 1. It could just as well have been 0 or -5.

1.7.3 Interval Scale

On interval measurement scales, one unit on the scale represents the same magnitude on the trait or characteristic being measured across the whole range of the scale. For example, if anxiety were measured on an interval scale, then a difference between a score of 10 and a score of 11 would represent the same difference in anxiety as would a difference between a score of 50 and a score of 51. Interval scales do not have a "true" zero point, however, and therefore it is not possible to make statements about how many times higher one score is than another. A good example of an interval scale is the Fahrenheit scale for temperature. Equal differences on this scale represent equal differences in temperature, but a temperature of 30 degrees is not twice as warm as one of 15 degrees.

1.7.4 Ratio Scale

Ratio scales are like interval scales except they have true zero points. A good example is the Kelvin scale of temperature. This scale has an absolute zero. Thus, a temperature of 300 Kelvin is twice as high as a temperature of 150 Kelvin.

Core text

S.P Gupta (2004): Introduction to statistical methods 23rd-ed: vikas publishing house New Delhi

2. Futher reading

• Saleemi N.A (1997), Statistics Simplified Reprinted January 2011: Nairobi, Saleemi Publication limited.

• Saleemi N.A (1992), Quantitative Techniques: Nairobi, Saleemi Publication limited.

LESSON 3

CHAPTER 3

COLLECTION OF DATA

For any statistical enquiry, the basic objective is to collect facts and figures relating to a particular phenomenon for further statistical analysis The process of counting, enumeration or measurement together with systematic recording of results is called collection of statistical data

Data types

2.1 Primary and Secondary Data

Primary data is data that you collect yourself using such methods as:

- questionnaires
- interviews
- focus group interviews
- observation
- case-studies
- diaries

- critical incidents

The primary data, which is generated by the above methods, may be qualitative in nature (usually in the form of words) or quantitative (usually in the form of numbers or where you can make counts of words used

2.2.1 Questionnaires

Questionnaires are a popular means of collecting data, but are difficult to design and often require many rewrites before an acceptable questionnaire is produced.

Advantages:

- Can be used as a method in its own right or as a basis for interviewing or a telephone survey.
- Can be posted, e-mailed or faxed.
- Can cover a large number of people or organisations.
- Wide geographic coverage.
- Relatively cheap.
- No prior arrangements are needed.
- Avoids embarrassment on the part of the respondent.
- Respondent can consider responses.
- Possible anonymity of respondent.
- No interviewer bias.

Disadvantages:

- Design problems.

- Questions have to be relatively simple.
- Historically low response rate (although inducements may help).
- Time delay whilst waiting for responses to be returned.
- Require a return deadline.
- Several reminders may be required.
- Assumes no literacy problems.
- No control over who completes it.
- Not possible to give assistance if required.
- Problems with incomplete questionnaires.
- Replies not spontaneous and independent of each other.
- Respondent can read all questions beforehand and then decide whether to complete or not. For example, perhaps because it is too long, too complex, uninteresting, or too personal

2.2.2 Interviews

Interviewing is a technique that is primarily used to gain an understanding of the underlying reasons and motivations for people's attitudes, preferences or behaviour. Interviews can be undertaken on a personal one-to-one basis or in a group. They can be conducted at work, at home, in the street or in a shopping centre, or some other agreed location.

Personal interview

Advantages:

- Serious approach by respondent resulting in accurate information.

- Good response rate.
- Completed and immediate.
- Possible in-depth questions.
- Interviewer in control and can give help if there is a problem.
- Can investigate motives and feelings.
- Can use recording equipment.
- Characteristics of respondent assessed – tone of voice, facial expression, hesitation, etc.
- Can use props.
- If one interviewer used, uniformity of approach.
- Used to pilot other methods.

Disadvantages:

- Need to set up interviews.
- Time consuming.
- Geographic limitations.
- Can be expensive.
- Normally need a set of questions.
- Respondent bias – tendency to please or impress, create false personal image, or end interview quickly.
- Embarrassment possible if personal questions.
- Transcription and analysis can present problems – subjectivity.

- If many interviewers, training required.

2.2.3 Case-studies

The term case-study usually refers to a fairly intensive examination of a single unit such as a person, a small group of people, or a single company. Case-studies involve measuring what is there and how it got there.

Iit is historical.

It can enable the researcher to explore, unravel and understand problems, issues and relationships.

It does not allow the researcher to argue that from one case-study the results, findings or theory developed apply to other similar case-studies. The case looked at may be unique and, therefore not representative of other instances

The case-study method has four steps:

- Determine the present situation.

- Gather background information about the past and key variables.

- Test hypotheses. The background information collected will have been analysed for possible hypotheses. In this step, specific evidence about each hypothesis can be gathered. This step aims to eliminate possibilities which conflict with the evidence collected and to gain confidence for the important hypotheses. The culmination of this step might be the development of an experimental design to test out more rigorously the hypotheses developed, or it might be to take action to remedy the problem.

- Take remedial action. The aim is to check that the hypotheses tested actually work out in practice. Some action, correction or improvement is made and a re-check carried out on the situation to see what effect the change has brought about.

The case-study enables rich information to be gathered from which potentially useful hypotheses can be generated.

- It can be a time-consuming process.
- It is inefficient in researching situations which are already well structured and where the important variables have been identified.
- They lack utility when attempting to reach rigorous conclusions or determining precise relationships between variables

2.2.4 Diaries

A diary is a way of gathering information about the way individuals spend their time on professional activities.

Diaries can record either quantitative or qualitative data, and in management research can provide information about work patterns and activities.

Advantages:

- Useful for collecting information from employees.
- Different writers compared and contrasted simultaneously.
- Allows the researcher freedom to move from one organization to another.
- Researcher not personally involved.
- Diaries can be used as a preliminary or basis for intensive interviewing.
- Used as an alternative to direct observation or where resources are limited.

Disadvantages:

- Subjects need to be clear about what they are being asked to do, why and what you plan to do with the data.

- Diarists need to be of a certain educational level.
- Some structure is necessary to give the diarist focus, for example, a list of headings.
- Encouragement and reassurance are needed as completing a diary is time-consuming and can be irritating after a while.
- Progress needs checking from time-to-time.
- Confidentiality is required as content may be critical.
- Analyses problems, so you need to consider how responses will be coded before the subjects start filling in diaries.

Secondary data is collected from external sources such as:

- TV, radio, internet
- magazines, newspapers
- reviews
- research articles
- stories told by people you know

Primary data is expensive and difficult to acquire, but it's trustworthy. Secondary data is cheap and easy to collect, but must be treated with caution.

Data Classification

The process of grouping raw data into different classes or sub classes according to some characteristics.

The collected data, also known as raw data or ungrouped data are always in an un organized form and need to be organized and presented in meaningful and readily comprehensible form in order to facilitate further statistical analysis. It is,

therefore, essential for an investigator to condense a mass of data into more and more comprehensible form.

Classification is the first step in tabulation

Objectives of Classification

- It condenses the mass of data in an easy form.
- It eliminates unnecessary details.
- It facilitates comparison and highlights the significant aspect of data.
- It enables one to get a mental picture of the information and helps in drawing inferences.
- It helps in the statistical treatment of the information collected.

Types of classification:

There are four basic types of classification

- **Chronological classification**- Based on order of time eg monthly , yearly etc

Its done In ascending order

- **Geographical classification**-According to geographical area or place

  e.g population per country

- **Qualitative classification**- on basis of same attribute or quality eg male and female
- **Quantitative classification**

classification of data according to some characteristics that can be measured such as

height, weight, etc., For example the students of a college may be classified according to weight

In this type of classification there are two elements

- The variable i.e. weight
- The frequency e.g. number of students in each class.

Tabulation:

This is the process of summarizing classified or grouped data in the form of a table so that it is easily understood and an investigator is quickly able to locate the desired information.

A table is a systematic arrangement of classified data in columns and rows. A statistical table makes it possible for the investigator to present a huge mass of data in a detailed and orderly form.

Advantages of Tabulation:

1. It simplifies complex data and the data presented are easily Understood
2. It facilitates comparison of related facts
3. It facilitates computation of various statistical measures like averages, dispersion, correlation etc.
4. It presents facts in minimum possible space and unnecessary repetitions and explanations are avoided.
5. Tabulated data are good for references and they make it easier to present the information in the form of graphs and diagrams.

An ideal table should consist of the following main parts:

- Table number
- Title of the table
- Captions or column headings
- Stubs or row designation
- Body of the table
- Footnotes
- Sources of data

DIAGRAMATIC AND GRAPHICAL REPRESENTATION

1.FREQUENCY DISTRIBUTION

It is simply a table in which the data are grouped into classes and the number of cases which fall in each class are recorded. It shows the frequency of occurrence of different values of a single Phenomenon

A frequency distribution is constructed for three main reasons:

1. To facilitate the analysis of data.

2. To estimate frequencies of the unknown population distribution from the distribution of sample data

3. To facilitate the computation of various statistical measures

Forms of frequency distributions

a) Discrete (or) Ungrouped frequency distribution:

In this form of distribution, the frequency refers to discrete value. The data are presented in a way that exact measurement of units are clearly indicated.

Each class is distinct and separate from the other class. Non-continuity from one class to another class exist.

Data as such facts like the number of rooms in a house, the number of companies registered in a country, the number of children in a family, etc.

How to prepare

- count the number of times a particular value is repeated- the frequency of that class.
- In order to facilitate counting prepare a column of tallies.
- In another column, place all possible values of variable from the lowest to the highest.
- Then put a bar (Vertical line) opposite the particular value to which it relates.
- To facilitate counting, blocks of five bars are prepared and some space is left in between each block.
- count the number of bars and get frequency.

Example 1:

In a survey of 40 families in a village, the number of children per family was recorded and the following data obtained.

1 ,0, 3, 2, 1, 5, 6, 2,

2 ,1 ,0 ,3 ,4 ,2 ,1 ,6

3 ,2, 1, 5, 3, 3, 2, 4

2 ,2 ,3 ,0 ,2 ,1 ,4 ,5

3 ,3, 4, 4, 1, 2, 4, 5

Represent the data in the form of a discrete frequency distribution.

Solution:

Frequency distribution of the number of children

Number of children	**Tally marks**	**frequency**
0	**111**	**3**
1	**~~1111~~ 11**	**7**
2	**~~1111~~ ~~1111~~**	**10**
3	**~~1111~~ 111**	**8**
4	**~~1111~~ 1**	**6**
5	**1111**	**4**
6	**11**	**2**
	total	**40**

b) Continuous frequency distribution:

In this form of distribution refers to groups of values. This becomes necessary in the case of some variables which can take any fractional value and in which case an exact measurement is not possible. Hence a discrete variable can be presented in the form of a continuous frequency distribution.

Wage distribution of 100 employees

Weekly wages (sh)	Number of workers
50-100	4
100-150	12
150-200	22
200-250	33
250-300	16
300-350	8
350-400	5
Total	100

Basic Terms

a) Class limits:

T he class limits are the lowest and the highest values that can be included in the class. For example, take the class 30-40.The lowest value of the class is 30 and highest class is 40. The two

boundaries of class are known as the lower limits and the upper limit of the class.

The lower limit of a class is the value below which there can be no item in the class.

The upper limit of a class is the value above which there can be no item to that class. Of the

class 60-79, 60 is the lower limit and 79 is the upper limit, i.e. in the case there can be no value which is less than 60 or more than

79. The way in which class limits are stated depends upon the nature of the data. In statistical calculations, lower class limit is denoted by L and upper class limit by U.

b) Class Interval:

The class interval may be defined as the size of each

grouping of data. For example, 50-75, 75-100, 100-125... are class

intervals. Each grouping begins with the lower limit of a class interval and ends at the lower limit of the next succeeding class interval

c) Width or size of the class interval:

The difference between the lower and upper class limits is called Width or size of class interval and is denoted by ' C' .

d) Range:

The difference between largest and smallest value of the observation is called The Range and is denoted by ' R' ie

R = Largest value – Smallest value

R = L - S

e) Mid-value or mid-point:

The central point of a class interval is called the mid value or mid-point. It is found out by adding the upper and lower limits of a class and dividing the sum by 2.

(i.e.) Midvalue = (L+ U)/2

For example, if the class interval is 20-30 then the mid-value is

(20 +30)/2

f) Frequency:

Number of observations falling within a particular class interval is called frequency of that class.

The total frequency indicate the total number of observations considered in a frequency distribution.

g) Number of class intervals:

The number of class interval should not be too many. For an ideal frequency distribution, the number of class intervals

To decide the number of class intervals for the frequency distributive in the whole data choose the lowest and the highest of the values. The difference between them will enable us to decide the class intervals.(use intuition)

2. Apply Sturges' Rule.

The rule states that the number of classes can be determined by the formula

K = 1 + 3. 322 log10 N

Where N = Total number of observations

log = logarithm of the number

K = Number of class intervals.

Thus if the number of observation is 10, then the number of class intervals is

K = 1 + 3. 322 log 10 = 4.322---4

If 100 observations are being studied, the number of class interval is

K = 1 + 3. 322 log 100 = 7.644-----8

h) Size of the class interval:

Since the size of the class interval is inversely proportional to the number of class interval in a given distribution. The approximate value of class width of the class interval ' C' is obtained by using sturges' rule as Size of class interval

C = Range/number of class intervals

=Range/1+3.322 log N

Where Range = Largest Value – smallest value in the distribution.

Types of class intervals:

There are three methods of classifying the data according to class intervals namely

- Exclusive method
- Inclusive method
- Open-end classes

a) Exclusive method:

In exclusive method, the class intervals are so fixed that the upper limit of one class is the lower limit of the next class.

This method ensures continuity of data

Its widely used in practice

Example

Expenditure	**No. of families**
0 – 5000	60
5000-10000	95
10000-15000	122
15000-20000	83
20000-25000	40
	Total 400

b) Inclusive method:

This method avoids the overlapping of class intervals

Both the lower and upper limits are included in the class interval.

This type of classification may be used for a grouped frequency distribution for discrete variable like members in a family, number of workers in a factory etc., where the variable may

take only integral values. It cannot be used with fractional values

like age, height, weight etc.

Class interval	**Frequency**
5- 9	7
10-14	12
15-19	15
20-29	21
30-34	10
35-39	5
Total	70

In case of continuous variables, the exclusive method must be used while the inclusive method should be used in case of discrete variable

c) Open end classes:

A class limit is missing either at the lower end of the first class interval or at the upper end of the last class interval or both are not specified. Its frequently used when there are few very high values or few very low values which are far apart from the majority of observations.

The example for the open-end classes as follows :

Salary Range	**No of workers**
Below 2000	7
2000 – 4000	5
4000 – 6000	6
6000 – 8000	4
8000 and above	3

Preparation of frequency table

The first step is to divide the observed range of variable into a suitable number of class-intervals and to record the number of observations in each class.

Example

The following data on weights of fifty college students. Construct a frequency table to represent the data

42 62 46 54 41 37 54 44 32 45
47 50 58 49 51 42 46 37 42 39
54 39 51 58 47 64 43 48 49 48
49 61 41 40 58 49 59 57 57 34
56 38 45 52 46 40 63 41 51 41

Apply sturges' rule

Number of class intervals: 1+3.322 log N=6.64 Approximately 7

Size of Class intervals = c = range/1+3.322 log N, (64 – 32)/1+3.322 log (50)

=32/6.64=4.8 Approx….five

The required frequency distribution is prepared using tally marks as given below:

Class Interval	**Tally marks**	**Frequency**
30-35		2
35-40		6
40-45		12
45-50		14
50-55		6
55-60		6
60-65		4
Total		50

2. Percentage frequency table

It is also called relative frequency table

The percentage frequency distribution facilitates easy comparability especially when the total number of items are large and highly different from one distribution to another. In percentage frequency table actual frequencies are converted into percentages. The percentages are calculated by using the formula given below:

Frequency percentage = Actual Frequency/Total Frequency× 100

An example is given below to construct a percentage frequency table.

Marks No. of students	Frequency	percentage
0-10	3	6
10-20	8	16
20-30	12	24
30-40	3	4
40-50	6	12
50-60	4	8
Total	50	100

3. Cumulative frequency table

Cumulative frequency distribution has a running total of the values. It is constructed by adding the frequency of the first class interval to the frequency of the second class interval. Again add

that total to the frequency in the third class interval continuing until the final total appearing opposite to the last class interval will be the total of all frequencies. The cumulative frequency may be downward or upward.

Example

Age (yrs)	No. men	Less than c.f	More than c.f
15-20	3	3	64
20-25	7	10	61
25-30	15	25	54
30-35	21	46	39

35-40	12	58	18
40-45	6	64	6

4. Histogram:

A histogram is a bar chart or graph showing the frequency of occurrence of each value of the variable being analysed. In histogram, data are plotted as a series of rectangles. Class intervals are shown on the 'X-axis' and the frequencies on the 'Y-axis' . The height of each rectangle represents the frequency of the class interval. Each rectangle is formed with the other so as to give a continuous picture.

Example

For the following data, draw a histogram.

Marks	Number of Students
21-30	6
31-40	15
41-50	22
51-60	31
61-70	17
71-80	9

Solution:

For drawing a histogram, the frequency distribution should be continuous. If it is not continuous, then first make it continuous as follows.

Marks	Number of Students
20.5-30.5	6
30.5-40.5	15
40.5-50.5	22
50.5-60.5	31
60.5-70.5	17
70.5-80.5	9

4. Frequency Polygon

If we mark the midpoints of the top horizontal sides of the rectangles in a histogram and join them by a straight line, the figure so formed is called a Frequency Polygon. This is done under the assumption that the frequencies in a class interval are evenly distributed throughout the class.

5 Frequency Curve

If the middle point of the upper boundaries of the rectangles of a histogram is corrected by a smooth freehand curve, then that diagram is called frequency curve. The curve should begin and end at the base line.

example

Draw a frequency curve for the following data.

Monthly Wages(sh.)	No. of family
0-1000	21
1000-2000	35
2000-3000	56
3000-4000	74
4000-5000	63
5000-6000	40
6000-7000	29
7000-8000	14

(see paper)

6.Ogive

This curve is obtained by plotting cumulative frequencies.

There are two methods of constructing ogive namely:

1. The ‘ less than ogive’ method
2. The ‘more than ogive’ method.

In less than ogive method we start with the upper limits of the classes and go adding the frequencies. When these frequencies are plotted, we get a rising curve. In more than

ogive method, we start with the lower limits of the classes and from the total frequencies we subtract the frequency of each class. When these frequencies are plotted we get a declining curve.

Example 15:

Draw the Ogives for the following data.

Class interval	Frequency	cf
20-30	4	4
30-40	6	10
40-50	13	23
50-60	25	48
60-70	32	80
70-80	19	99
80-90	8	107
90-100	3	110

Solution:

Class limit	Less than ogive	More than ogive
20	0	110
30	4	106
40	10	100
50	23	87
60	48	62
70	80	30
80	99	11
90	107	3
100	110	0

MEASURES OF CENTRAL TENDENCY

A measure of central tendency is a representative number that summarises the whole data set. Its also know as an average or a measure of locations

E.G

Mean, Median and mode.....simple averages

Harmonic and geometric mean.....special averages

Characteristics for a good or an ideal average:

The following properties should possess for an ideal average.

1. It should be rigidly defined.
2. It should be easy to understand and compute.
3. It should be based on all items in the data.
4. Its definition shall be in the form of a mathematical formula.
5. It should be capable of further algebraic treatment.
6. It should have sampling stability.
7. It should be capable of being used in further statistical computations or processing

Arithmetic mean or mean:

Arithmetic mean or simply the mean of a variable is defined as the sum of the observations divided by the number of observations. If the variable x assumes n values x1, x2 ... xn then the

mean, $\bar{x}$, is given by $\frac{i}{n}\sum_{i=1}^{n} x_i$ (For ungrouped or raw data)

For grouped Data :

The mean for grouped data is obtained from the following formula:

$$\bar{x} = \frac{\sum fx}{N}$$

where x = the mid-point of individual class

f = the frequency of individual class

N = the sum of the frequencies or total frequencies.

Example *:*

Following is the distribution of persons according to different income groups. Calculate arithmetic mean. (see HR paper)

Income sh (1000)

0-10 10-20 20-30 30-40 40-50 50-60 60-70

6 8 10 12 7 4 3

Merits and demerits of Arithmetic mean

- **Merits**

1. It is rigidly defined.
2. It is easy to understand and easy to calculate
3. If the number of items is sufficiently large, it is more accurate and more reliable.
4. It is a calculated value and is not based on its position in the series
5. It is possible to calculate even if some of the details of the data are lacking.
6. Of all averages, it is affected least by fluctuations of sampling.
7. It provides a good basis for comparison.

- **Demerits:**

1. It cannot be obtained by inspection nor located through a frequency graph.
2. It cannot be in the study of qualitative phenomena not capable of numerical measurement i.e. Intelligence, beauty, honesty etc.,
3. It can ignore any single item only at the risk of losing its accuracy.
4. It is affected very much by extreme values.
5. It cannot be calculated for open-end classes.
6. It may lead to fallacious conclusions, if the details of the data from which it is computed are not given

Median

The median is that value which divides the data group into two equal parts, one part comprising all values greater, and the other, all values less than median.

Ungrouped or Raw data :

Arrange the given values in the increasing or decreasing order. If the number of values are odd, median is the middle value. If the number of values are even, median is the mean of middle two values.

Example

When odd number of values are given. Find median for the following data

25, 18, 27, 10, 8, 30, 42, 20, 53

Solution:

Arranging the data in the increasing order 8, 10, 18, 20, 25,27, 30, 42, 53

The middle value is the 5th item i.e., 25 is the median. W hen even number of values is given.

Example

Find median for the following data

5, 8, 12, 30, 18, 10, 2, 22

Solution:

Arranging the data in the increasing order 2, 5, 8, 10, 12,18, 22, 30.Here median is the mean of the middle two items (i.e) mean of (10,12) i.e.

= 10 +12 median = 11

Grouped Data:

In a grouped distribution, values are associated with frequencies. Grouping can be in the form of a discrete frequency distribution or a continuous frequency distribution. Whatever may be the type of distribution , cumulative frequencies have to be calculated to know the total number of items.

Discrete Series:

Step1: Find cumulative frequencies.

Step2: Find

$$\left[\frac{N+1}{2}\right]$$

Step3: See in the cumulative frequencies the value just greater than

$$\left[\frac{N+1}{2}\right]$$

Step4: Then the corresponding value of x is median.

Example :

The following data pertaining to the number of members in a family. Find median size of the family.

Number of members **x**	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	1	3	5	6	10	13	9	5	3	2	2	1

Solution:

X	f	cf
1	1	1
2	3	4
3	5	9
4	6	15
5	10	25
6	13	38
7	9	47
8	5	52
9	3	55
10	2	57
11	2	59
12	1	60
	60	

$\left[\frac{60+1}{2}\right] = 30.5$

The cumulative frequencies just greater than 30.5 is 38.and the value of x corresponding to 38 is 6. Hence the median size is 6 members per family

Continuous Series:

The steps given below are followed for the calculation of median in continuous series.

Step1: Find cumulative frequencies.

Step2: Find

N/2

Step3: See in the cumulative frequency the value first greater than N/2. Then the corresponding class interval is called the Median class. Then apply the formula

Median = $l + \frac{N/2 - m}{f} \times c$ Where l = Lower limit of the median class, m = cumulative frequency preceding the median class, c = width of the median class, f =frequency in the median class. N=Total frequency. If the class intervals are given in inclusive type convert them into exclusive type and call it as true class interval and consider lower limit in this case.

Example 7:

Determine the median of the data in the table below using Formula method

IQ	No of residents
0 – 20	6
20 – 40	18
40 – 60	32
60 – 80	48
80 – 100	27
100 – 120	13
120 – 140	2

$$= 60 + \frac{73.5 - 56}{48} \times 20$$

$$= 60 + 7.29$$

$$= 67.29$$

Merits of Median :

1. Median is not influenced by extreme values because it is a positional average.
2. Median can be calculated in case of distribution with open end intervals.
3. Median can be located even if the data are incomplete.
4. Median can be located even for qualitative factors such as ability, honesty etc.

Demerits of Median:

1. A slight change in the series may bring drastic change in median value.
2. In case of even number of items or continuous series, median is an estimated value other than any value in the series.
3. It is not suitable for further mathematical treatment except its use in mean deviation.
4. It is not taken into account all the observations

Quartiles :

The quartiles divide the distribution in four parts. There are three quartiles. The second quartile divides the distribution into two halves and therefore is the same as the median. The first (lower) quartile (Q1) marks off the first one-fourth, the third (upper) quartile (Q3) marks off the three-fourth. First arrange the given data in the increasing order and use the formula for Q1 and Q3

$Q_1 = \frac{n+1}{4}$ th item, $Q_3 = \frac{n+1}{4} \times 3$ th item

$Q.D = \frac{Q_3 - Q_1}{2}$ Quartile Deviation

Example 22 :

Compute quartiles for the data given below 25,18,30, 8, 15,

5, 10, 35, 40, 45

Solution

5, 8, 10, 15, 18,25, 30,35,40, 45

Q1=10+1/4=2.75th item, $8 + \frac{10-8}{}.\frac{3}{4}$=9. 5

Q3=3(2.75) = 8.25th item= 35+1/4(40-35) =36.25

Discrete Series

Step1: Find cumulative frequencies.

Step2: Find $\frac{N+1}{4}$

Step3: See in the cumulative frequencies, the value just greater than

$\frac{N+1}{4}$, then the corresponding value of x is Q1

Step4: Find $\frac{3}{4}\left[\frac{N+1}{4}\right]$

Step5: See in the cumulative frequencies, the value just greater than $\frac{3}{4}\left[\frac{N+1}{4}\right]$, then the corresponding value of x is Q3

Example 23:

Compute quartiles for the data given bellow.

x f c.f

5	4	4
8	3	7
12	2	9
15	4	13
19	5	18
24	2	20
30	4	24
	Total 24	

Solution

Q1= 24+1/4= 6.25th item = 8

Q3=3/4(6.25) = 18.75th =24

Continuous series

Step1: Find cumulative frequencies

Step2: Find N/4

Step3: See in the cumulative frequencies, the value just greater than N/4, then the corresponding class interval is called first quartile class. Find 3/4N. See in the cumulative frequencies the value

just greater than ¾(N) then the corresponding class interval is called 3rd quartile class.

Then apply the respective formulae $Q_1 = l_1 + \frac{N/4 - m_1}{f1} \times c1$ and $Q_3 = l_3 + \frac{3N/4 - m_3}{f3} \times c3$

Where $l1$ = lower limit of the first quartile class

$f1$ = frequency of the first quartile class

$c1$ = width of the first quartile class

$m1$ = c.f. preceding the first quartile class

$l3$ = 1ower limit of the 3rd quartile class

$f3$ = frequency of the 3rd quartile class

$c3$ = width of the 3rd quartile class

$m3$ = c.f. preceding the 3rd quartile class

Example

C . I.	f	cf
0-10	11	11
10-20	18	29
20-30	25	54
30-40	28	82
40-50	30	112
50-60	33	145
60-70	22	167
70-80	15	182
80-90	12	194
90-100	10	204
	204	

N/4=204/4=51, 3(204/4) = 153

MODE

The mode refers to that value in a distribution, which occur most frequently. It is an actual value, which has the highest concentration of items in and around it. It shows the centre of concentration of the frequency in around a given value. Therefore, where the purpose is to know the point of the highest concentration it is preferred. It is, thus, a positional measure.

Computation of the mode:

Ungrouped or Raw Data:

For ungrouped data or a series of individual observations, mode is often found by mere inspection.

Example

2 , 7, 10, 15, 10, 17, 8, 10, 2\ Mode = M0=10

In some cases the mode may be absent while in some cases there may be more than one mode.

Example

12, 10, 15, 24, 30 (no mode)

7, 10, 15, 12, 7, 14, 24, 10, 7, 20, 10\ the modes are 7 and 10

Grouped Data

For Discrete distribution, see the highest frequency and corresponding value of X is mode.

Continuous distribution

See the highest frequency then the corresponding value of class interval is called the modal class. Then apply the formula

$$m_0 = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times c$$

l = Lower limit of the model class ,Δ1 = f1-f0, Δ2 =f1-f2 ,f1 = frequency of the modal class,f0 = frequency of the class preceding the modal class,f2 = frequency of the class succeeding the modal class or simply

$$m_0 = l + \frac{f_1 - f_0}{2f_1 - f_0 - f_2}$$

Example

Calculate mode for the following :

C- I	f
0-50	5
50-100	14
100-150	40

150-200	91
200-250	150
250-300	87
300-350	60
350-400	38
400 and above	15

Solution

The highest frequency is 150 and corresponding class interval is

200 – 250, which is the modal class.

Here l=200,f1=150,f0=91, f2=87, C=50

Mode=224.18

Merits of Mode:

1. It is easy to calculate and in some cases it can be located mere inspection
2. Mode is not at all affected by extreme values.
3. It can be calculated for open-end classes.
4. It is usually an actual value of an important part of the series.
5. In some circumstances it is the best representative of data.

Demerits of mode:

1. It is not based on all observations.
2. It is not capable of further mathematical treatment.
3. Mode is ill-defined generally, it is not possible to find mode in some cases.
4. As compared with mean, mode is affected to a great extent, by sampling fluctuations.
5. It is unsuitable in cases where relative importance of items has to be considered.

Core text

S.P Gupta (2004): Introduction to statistical methods 23rd-ed: vikas publishing house New Delhi

2. Futher reading

• Saleemi N.A (1997), Statistics Simplified Reprinted January 2011: Nairobi, Saleemi Publication limited.

• Saleemi N.A (1992), Quantitative Techniques: Nairobi, Saleemi Publication limited.

LESSON 4

MEASURES OF DISPERSION

6.1 INTRODUCTION TO MEASURES OF DISPERSION

The measures of dispersion are very useful in statistical work because they indicate whether the

rest of the data are scattered around the mean or away from the mean.

If the data is approximately dispersed around the mean then the measure of dispersion obtained

will be small therefore indicating that the mean is a good representative of the sample data. But

on the other hand, if the figures are not closely located to the mean then the measures of

dispersion obtained will be relatively big indicating that the mean does not represent the data

sufficiently

The commonly used measures of dispersion are

i. The range

ii. The absolute mean deviation

iii. The standard deviation

iv. The semi – interquartile and quartile deviation

v. The 10th and 90th percentile range

vi. Variance

6.2 RANGE

The range is defined as the difference between the highest and the smallest values in a frequency

distribution. This measure is not very efficient because it utilizes only 2 values in a given

frequency distribution. However the smaller the value of the range, the less dispersed the

observations are from the arithmetic mean and vice versa

Example 1:

The following are the prices of 4 kgs of beans in Mathare slums market

Day	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday
Price(Ksh.)	200	210	208	160	200	250

Required:

Find the range and Co-efficient of range.

Solution

Range=L-S

250-160=90

Co-efficient of Range=L-S/L+S

=250-160/250+160=90/410

=0.22

INTERQUARTILE RANGE

This is a measure of dispersion which involves the use of quartile. A quartile is a mark or a value

which lies at the boundary of a division when any given set of data is divided into four equal

divisions.

Each of such divisions normally carries 25% of all the observations

The semi interquartile range is a good measure of dispersion because it shows how the rest of the

data are generally spread around the mean

The quartiles normally used are three namely;

i. The lower quartile (first quartile Q1) this usually binds the lower 25% of the data

ii. The median (second quartile Q2)

iii. The upper quartile (third quartile Q3)

The semi-interquartile range,

Q3 - Q1

SIR =

2

Example 2:

The weights of 15 parcels recorded at the GPO were as follows:

16.2, 17, 20, 25(Q1) 29, 32.2, 35.8, 36.8(Q2) 40, 41, 42, 44(Q3) 49, 52, 55 (in kgs)

Required

Determine the semi interquartile range for the above data

$\underline{(Q3 - Q1}\ )/2$

=(44 - 25 19)/2= 8.5

Mean Deviation and Coefficient of Mean Deviation:

Mean Deviation:

The range and quartile deviation are not based on all observations. They are positional measures of dispersion. They do not show any scatter of the observations from an average. The mean deviation is measure of dispersion based on all items in a distribution.

Definition:

Mean deviation is the arithmetic mean of the deviations of a series computed from any measure of central tendency; i.e., the mean, median or mode, all the deviations are taken as positive i.e., signs are ignored. According to Clark and

Schekade, “Average deviation is the average amount scatter of the items in a distribution from either the mean or the median, ignoring the signs of the deviations”.

We usually compute mean deviation about any one of the three averages mean, median or mode. Some times mode may be ill defined and as such mean deviation is computed

from mean and median. Median is preferred as a choice between mean and median. But in general practice and due to wide applications of mean, the mean deviation is generally computed from mean. M.D can be used to denote mean deviation.

Coefficient of mean deviation:

Mean deviation calculated by any measure of central tendency is an absolute measure. For the purpose of comparing variation among different series, a relative mean deviation is required. The relative mean deviation is obtained by dividing the mean deviation by the average used for calculating mean deviation.

$$\text{Coefficient of mean deviation: } = \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}}$$

If the result is desired in percentage, the coefficient of mean

$$\text{deviation} = \frac{\text{Mean deviation}}{\text{Mean or Median or Mode}} \times 100$$

Computation of mean deviation – Individual Series:

1. Calculate the average mean, median or mode of the series.

2. Take the deviations of items from average ignoring signs and denote these deviations by |D|.

3. Compute the total of these deviations, i.e., S |D|

4. Divide this total obtained by the number of items.

$$\text{Symbolically: M.D.} = \frac{|D|}{N}$$

Calculate mean deviation from mean and median for the following data:

100,150,200,250,360,490,500,600,671 also calculate coefficients

of M.D.

Solution:

Mean = **369**

$Md = \sum |D|/n$

$=1570 / 9 = 174.44$

Mean deviation-Continuous series:

The method of calculating mean deviation in a continuous series same as the discrete series. In continuous series we have to find out the mid points of the various classes and take deviation of these points from the average selected. Thus

$$M.D = \frac{\sum f | D |}{N}$$

Where D = m - average

M = Mid point

Find out the mean deviation from mean from the following series.

Age in years	No.of persons
0-10	20
10-20	25
20-30	32
30-40	40
40-50	42
50-60	35
60-70	10
70-80	8

Solution:

Mean = 35

$$MD = \frac{\sum f|D|}{N} = 3193/212 = 15.06$$

Merits and Demerits of M.D :

Merits:

1. It is simple to understand and easy to compute.

2. It is rigidly defined.

3. It is based on all items of the series.

4. It is not much affected by the fluctuations of sampling.

5. It is less affected by the extreme items.

6. It is flexible, because it can be calculated from any average.

7. It is better measure of comparison.

Demerits:

1. It is not a very accurate measure of dispersion.

2. It is not suitable for further mathematical calculation.

3. It is rarely used. It is not as popular as standard deviation.

4. Algebraic positive and negative signs are ignored. It is mathematically unsound and illogical.

Standard Deviation:

Karl Pearson introduced the concept of standard deviation in 1893. It is the most important measure of dispersion and is widely used in many statistical formulae. Standard deviation is also

called Root-Mean Square Deviation. The reason is that it is the square–root of the mean of the squared deviation from the arithmetic mean. It provides accurate result. Square of standard deviation is called Variance.

Definition:

It is defined as the positive square-root of the arithmetic mean of the Square of the deviations of the given observation from their arithmetic mean. The standard deviation is denoted by the Greek letter σ (sigma)