AMBO UNIVERSITY WOLISO CAMPUS

INTRODUCTION TO STATISTICS (Stat2131)

DEPARTMENT OF BASIC COURSES

KEDIR BEKERU[1] (MSc. and MPP)

[1] Lecturer of Biostatistics, Department of Basic courses, Ambo University Woliso Campus
Email: kedirb@kedis.ac.kr

INTRODUCTION TO STATISTICS (Stat2131)

COURSE DESCRIPTION

Meaning of statistics, methods of data collection, methods of data presentation, measure of central tendency; measure of variation; moments, skewness, kurtosis; concepts of probability; counting techniques, probability distribution: Binomial, Poisson, Normal, t and Chi-square; sampling.

OBJECTIVES

- To introduce to the students to basic statistical knowledge on data collection and presentation methods, measure of central tendency, measure of variation and sampling technique.
- To demonstrate the importance and usefulness of statistics in real life and real data,
- To show how to present data informatively and clearly,
- To build up interest in statistics and hence encourage the students to study the subject further.

LEARNING OUTCOMES

At the end of the course students expected to

- Have a broad knowledge of the basic understanding statistical techniques demonstrated through principles of data collection, descriptive statistics, data analysis and sampling.
- Identify different techniques of sampling and understand the methods of data collection, organization, presentation, analyze and interpretation.
- Differentiate among common types of data and summarize and display them appropriately
- Learn some desirable properties of averages and measure of variation.
- Have basic skills in explanatory data analysis and problems solving,

COURSE OUTLINE

1. **Introduction**
   - 1.1 History, definitions and classification of statistics
   - 1.2 Stages in statistical investigation,
   - 1.3 Definition of some basic terms
   - 1.4 Application, uses and limitation of statistics
   - 1.5 Scales of measurements
2. **Methods of data presentation**
   - 2.1 Introduction
   - 2.2 Frequency distribution, qualitative, quantitative (absolute, relative, percentage cumulative
   - 2.3 Diagrammatic presentation of data: bar chart, pie chart, pictogram, steam and leaf plot
   - 2.4 Graphical presentation of data: histogram, frequency polygon, Ogive.
3. **Measure of central tendency**
   - 3.1 Introduction
     - 3.1.1 Objectives of measuring central tendency
     - 3.1.2 The summation notation
     - 3.1.3 Important characteristics of measure of central tendency
     - 3.1.4 Types of measures of central tendency
       - 3.1.4.1 The mean (Arithmetic, weighted, geometric and harmonic)
       - 3.1.4.2 The mode
       - 3.1.4.3 The median
       - 3.1.4.4 The quintiles (quartiles, deciles, percentiles)
4. **Measure of variation**
   - 4.1 Introduction
   - 4.2 Objectives of measuring variation
   - 4.3 Absolute and relative variation
   - 4.4 Types of measure of variation

CHAPTER 1

1. INTRODUCTION

Definition and classifications of statistics

Definition:

We can define statistics in two ways.

1. Plural sense (lay man definition).
   It is an aggregate or collection of numerical facts.
2. Singular sense (formal definition)
   Statistics is defined as the science of collecting, organizing, presenting, analyzing and interpreting numerical data for the purpose of assisting in making a more effective decision.

Classifications:

Depending on how data can be used statistics is sometimes divided in to two main areas or branches.

1. **Descriptive Statistics**: is concerned with summary calculations, graphs, charts and tables.
2. **Inferential Statistics**: is a method used to generalize from a sample to a population. For example, the average income of all families (the population) in Ethiopia can be estimated from figures obtained from a few hundred (the sample) families.

- It is important because statistical data usually arises from sample.
- Statistical techniques based on probability theory are required**.**

Stages in Statistical Investigation

There are five stages or steps in any statistical investigation.

1. **Collection of data**: the process of measuring, gathering, assembling the raw data up on which the statistical investigation is to be based.
   - Data can be collected in a variety of ways; one of the most common methods is through the use of survey. Survey can also be done in different methods, three of the most common methods are:
     - Telephone survey
     - Mailed questionnaire
     - Personal interview.

   Exercise: discuss the advantage and disadvantage of the above three methods with respect to each other.

2. **Organization of data**: Summarization of data in some meaningful way, e.g. table form
3. **Presentation of the data**: The process of re-organization, classification, compilation, and summarization of data to present it in a meaningful form.
4. **Analysis of data**: The process of extracting relevant information from the summarized data, mainly through the use of elementary mathematical operation.
5. **Inference of data**: The interpretation and further observation of the various statistical measures through the analysis of the data by implementing those methods by which conclusions are formed and inferences made.
    - Statistical techniques based on probability theory are required.

Definitions of some terms

a. **Statistical Population**: It is the collection of all possible observations of a specified characteristic of interest (possessing certain common property) and being under study. An example is all of the students in AAU 3101 course in this term.
b. **Sample**: It is a subset of the population, selected using some sampling technique in such a way that they represent the population.
c. **Sampling:** The process or method of sample selection from the population.
d. **Sample size:** The number of elements or observation to be included in the sample.
e. **Census:** Complete enumeration or observation of the elements of the population. Or it is the collection of data from every element in a population
f. **Parameter:** Characteristic or measure obtained from a population.
g. **Statistic:** Characteristic or measure obtained from a sample.
h. **Variable:** It is an item of interest that can take on many different numerical values.

Types of Variables or Data:

1. **Qualitative Variables** are nonnumeric variables and can't be measured. Examples include gender, religious affiliation, and state of birth.
2. **Quantitative Variables** are numerical variables and can be measured. Examples include balance in checking account, number of children in family. Note that quantitative variables are either discrete (which can assume only certain values, and there are usually "gaps" between the values, such as the number of bedrooms in your house) or continuous (which can assume any value within a specific range, such as the air pressure in a tire.)

Applications, Uses and Limitations of statistics

Applications of statistics:

- In almost all fields of human endeavor.
- Almost all human beings in their daily life are subjected to obtaining numerical facts e.g. abut price.
- Applicable in some process e.g. invention of certain drugs, extent of environmental pollution.
- In industries especially in quality control area.

Uses of statistics:

The main function of statistics is to enlarge our knowledge of complex phenomena. The following are some uses of statistics:

1. It presents facts in a definite and precise form.
2. Data reduction.
3. Measuring the magnitude of variations in data.
4. furnishes a technique of **comparison**
5. Estimating unknown population characteristics.
6. Testing and formulating of hypothesis.
7. Studying the relationship between two or more variable.
8. Forecasting future events.

Limitations of statistics

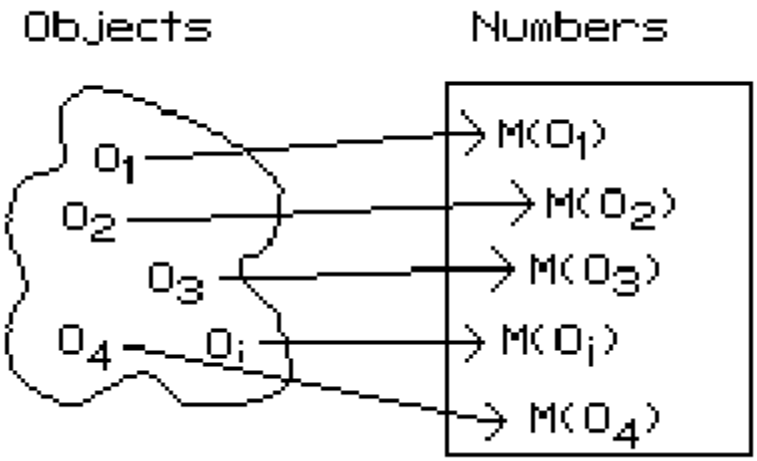
As a science statistics has its own limitations. The following are some of the limitations:

- Deals with mainly quantitative information.
- Deals with only aggregate of facts and not with individual data items.
- Statistical data are only approximately and not mathematical correct.
- Statistics can be easily misused and therefore should be used be experts.

Scales of measurement

Proper knowledge about the nature and type of data to be dealt with is essential in order to specify and apply the proper statistical method for their analysis and inferences. Measurement scale refers to the property of value assigned to the data based on the properties of order, distance and fixed zero.

In mathematical terms measurement is a functional mapping from the set of objects $\{O_i\}$ to the set of real numbers $\{M(O_i)\}$.



The goal of measurement systems is to structure the rule for assigning numbers to objects in such a way that the relationship between the objects is preserved in the numbers assigned to the objects. The different kinds of relationships preserved are called properties of the measurement system.

Order

The property of order exists when an object that has more of the attribute than another object, is given a bigger number by the rule system. This relationship must hold for all objects in the "real world".

The property of ORDER exists

When for all i, j if $O_i > O_j$, then $M(O_i) > M(O_j)$.

Distance

The property of distance is concerned with the relationship of differences between objects. If a measurement system possesses the property of distance it means that the unit of measurement means the same thing throughout the scale of numbers. That is, an inch is an inch, no matters were it falls - immediately ahead or a mile downs the road.

More precisely, an equal difference between two numbers reflects an equal difference in the "real world" between the objects that were assigned the numbers. In order to define the property of distance in the mathematical notation, four objects are required: O_i, O_j, O_k, and O_l . The difference between objects is represented by the "-" sign; $O_i - O_j$ refers to the actual "real world" difference between object i and object j, while $M(O_i) - M(O_j)$ refers to differences between numbers.

The property of DISTANCE exists, for all i, j, k, l

If $O_i - O_j \geq O_k - O_l$ then $M(O_i) - M(O_j) \geq M(O_k) - M(O_l)$.

Fixed Zero

A measurement system possesses a rational zero (fixed zero) if an object that has none of the attribute in question is assigned the number zero by the system of rules. The object does not need to really exist in the "real world", as it is somewhat difficult to visualize a "man with no height". The requirement for a rational zero is this: if objects with none of the attribute did exist would they be given the value zero. Defining O_0 as the object with none of the attribute in question, the definition of a rational zero becomes:

The property of FIXED ZERO exists if $M(O_0) = 0$.

The property of fixed zero is necessary for ratios between numbers to be meaningful.

SCALE TYPES

Measurement is the assignment of numbers to objects or events in a systematic fashion. Four levels of measurement scales are commonly distinguished: nominal, ordinal, interval, and ratio and each possessed different properties of measurement systems.

Nominal Scales

Nominal scales are measurement systems that possess none of the three properties stated above.

- Level of measurement which classifies data into mutually exclusive, all inclusive categories in which no order or ranking can be imposed on the data.
- No arithmetic and relational operation can be applied.

Examples:

- Political party preference (Republican, Democrat, or Other,)
- Sex (Male or Female.)
- Marital status(married, single, widow, divorce)
- Country code
- Regional differentiation of Ethiopia.

Ordinal Scales

Ordinal Scales are measurement systems that possess the property of order, but not the property of distance. The property of fixed zero is not important if the property of distance is not satisfied.

- Level of measurement which classifies data into categories that can be ranked. Differences between the ranks do not exist**.**
- Arithmetic operations are not applicable but relational operations are applicable.
- Ordering is the sole property of ordinal scale.

Examples:

- Letter grades (A, B, C, D, F).
- Rating scales (Excellent, Very good, Good, Fair, poor).
- Military status.

Interval Scales

Interval scales are measurement systems that possess the properties of Order and distance, but not the property of fixed zero.

- Level of measurement which classifies data that can be ranked and differences are meaningful. However, there is no meaningful zero, so ratios are meaningless.
- All arithmetic operations except division are applicable.
- Relational operations are also possible.

Examples:

- IQ
- Temperature in o_F.

Ratio Scales

Ratio scales are measurement systems that possess all three properties: order, distance, and fixed zero. The added power of a fixed zero allows ratios of numbers to be meaningfully interpreted; i.e. the ratio of Bekele's height to Martha's height is 1.32, whereas this is not possible with interval scales.

- Level of measurement which classifies data that can be ranked, differences are meaningful, and there is a true zero. True ratios exist between the different units of measure.

- All arithmetic and relational operations are applicable.

Examples:

  - Weight
  - Height
  - Number of students
  - Age

The following present a list of different attributes and rules for assigning numbers to objects. Try to classify the different measurement systems into one of the four types of scales. (Exercise)

1. Your checking account number as a name for your account.
2. Your checking account balance as a measure of the amount of money you have in that account.
3. The order in which you were eliminated in a spelling bee as a measure of your spelling ability.
4. Your score on the first statistics test as a measure of your knowledge of statistics.
5. Your score on an individual intelligence test as a measure of your intelligence.
6. The distance around your forehead measured with a tape measure as a measure of your intelligence.
7. A response to the statement "Abortion is a woman's right" where "Strongly Disagree" = 1, "Disagree" = 2, "No Opinion" = 3, "Agree" = 4, and "Strongly Agree" = 5, as a measure of attitude toward abortion.
8. Times for swimmers to complete a 50-meter race
9. Months of the year Meskerm, Tikimit…
10. Socioeconomic status of a family when classified as low, middle and upper classes.
11. Blood type of individuals, A, B, AB and O.
12. Pollen counts provided as numbers between 1 and 10 where 1 implies there is almost no pollen and 10 that it is rampant, but for which the values do not represent an actual counts of grains of pollen.
13. Regions numbers of Ethiopia (1, 2, 3 etc.)
14. The number of students in a college;
15. The net wages of a group of workers;
16. the height of the men in the same town;

CHAPTER 2

2. METHODS OF DATA COLLECTION AND PRESNTATION

2.1 Introduction:

Data:

- What is data? Raw data is recorded information in its original collected form, whether it be counts or measurements.
- The raw material for statistics

- It can be obtained from:
  - ✓ Routinely kept records on book
  - ✓ Surveys
  - ✓ Census
  - ✓ Vital registration
  - ✓ Reports

2.2 Methods of Data Collection and tool

There are two sources of data:

1. **Primary Data**
   - Data measured or collect by the investigator or the user directly from the source.
   - Two activities involved: planning and measuring.

a) Planning:
   - Identify source and elements of the data.
   - Decide whether to consider sample or census.
   - If sampling is preferred, decide on sample size, selection method… etc.
   - Decide measurement procedure.
   - Set up the necessary organizational structure.

b) Measuring: there are different options.
   - Focus Group
   - Telephone Interview
   - Mail Questionnaires
   - Door-to-Door Survey
   - Mall Intercept
   - New Product Registration
   - Personal Interview and
   - Experiments are some of the sources for collecting the primary data.

2. **Secondary Data**
   - Data gathered or compiled from published and unpublished sources or files.
   - When our source is secondary data check that:

- The type and objective of the situations.
- The purpose for which the data are collected and compatible with the present problem.
- The nature and classification of data is appropriate to our problem.
- There are no biases and misreporting in the published data.

⇒ **The following are some of the methods used for collecting the data:**

- ▶ Observation
- ▶ Face-to-face and self-administered interviews
- ▶ Postal or mail method and telephone interviews
- ▶ Focus group discussions
- ▶ Desk review
- ▶ Experiments.
- ▶ Others' such as, life histories, Case studies, Nominal group techniques, etc.

Note: Data which are primary for one may be secondary for the other.

Common Problems in Data Collection

♥ The following are some of the common problems during data collection:

- ✓ Language barriers
- ✓ Lack of adequate time
- ✓ Expense
- ✓ Inadequately trained and experienced staff
- ✓ Invasion of privacy
- ✓ Bias
- ✓ Cultural norms

METHOD OF DATA PRESENTATION

Having collected and edited the data, the next important step is to organize it. That is to present it in a readily comprehensible condensed form that aids in order to draw inferences from it. It is also necessary that the like be separated from the unlike ones.

The presentation of data is broadly classified in to the following two categories:

- Tabular presentation
- Diagrammatic and Graphic presentation.

The process of arranging data in to classes or categories according to similarities technically is called *classification*.

Classification is a preliminary and it prepares the ground for proper presentation of data.

Definitions:

- Raw data: recorded information in its original collected form, whether it be counts or measurements, is referred to as raw data.
- Frequency: is the number of values in a specific class of the distribution.
- Frequency distribution: is the organization of raw data in table form using classes and frequencies.

There are three basic types of frequency distributions

- Categorical frequency distribution
- Ungrouped frequency distribution
- Grouped frequency distribution

There are specific procedures for constructing each type.

1) Categorical frequency Distribution:

Used for data that can be place in specific categories such as nominal, or ordinal. e.g. marital status.

Example: a social worker collected the following data on marital status for 25 persons. (M=married, S=single, W=widowed, D=divorced)

M	S	D	W	D
S	S	M	M	M
W	D	S	M	M
W	D	D	S	S
S	W	W	D	D

Solution:

Since the data are categorical, discrete classes can be used. There are four types of marital status M, S, D, and W. These types will be used as class for the distribution. We follow procedure to construct the frequency distribution.

Step 1: Make a table as shown.

Class	Tally	Frequency	Percent
(1)	(2)	(3)	(4)
M			
S			
D			
W			

Step 2: Tally the data and place the result in column (2).

Step 3: Count the tally and place the result in column (3).

Step 4: Find the percentages of values in each class by using;

$\% = \frac{f}{n} * 100$ Where f= frequency of the class, n=total number of value.

Percentages are not normally a part of frequency distribution but they can be added since they are used in certain types diagrammatic such as pie charts.

Step 5: Find the total for column (3) and (4).

Combing all the steps one can construct the following frequency distribution.

Class	Tally	Frequency	Percent
(1)	(2)	(3)	(4)
M	///	5	20
S	/// //	7	28
D	/// //	7	28
W	/// /	6	24

2) Ungrouped frequency Distribution:

-Is a table of all the potential raw score values that could possible occur in the data along with the number of times each actually occurred.

-Is often constructed for small set or data on discrete variable.

Constructing ungrouped frequency distribution:

- First find the smallest and largest raw score in the collected data.
- Arrange the data in order of magnitude and count the frequency.
- To facilitate counting one may include a column of tallies.

Example:

The following data represent the mark of 20 students.

80	76	90	85	80
70	60	62	70	85
65	60	63	74	75
76	70	70	80	85

Construct a frequency distribution, which is ungrouped.

Solution:

Step 1: Find the range, Range=Max-Min=90-60=30.
Step 2: Make a table as shown
Step 3: Tally the data.
Step 4: Compute the frequency.

Mark	Tally	Frequency
60	//	2
62	/	1
63	/	1
65	/	1
70	////	4
74	/	1
75	//	2
76	/	1
80	///	3
85	///	3
90	/	1

Each individual value is presented separately, that is why it is named ungrouped frequency distribution.

3) Grouped frequency Distribution:

When the range of the data is large, the data must be grouped in to classes that are more than one unit in width.

Definitions:

- **Grouped Frequency Distribution:** a frequency distribution when several numbers are grouped in one class.
- **Class limits:** Separates one class in a grouped frequency distribution from another. The limits could actually appear in the data and have gaps between the upper limits of one class and lower limit of the next.
- **Units of measurement (U):** the distance between two possible consecutive measures. It is usually taken as 1, 0.1, 0.01, 0.001, -----.
- **Class boundaries:** Separates one class in a grouped frequency distribution from another. The boundaries have one more decimal places than the row data and therefore do not appear in the data. There is no gap between the upper boundary of one class and lower boundary of the next class. The lower class boundary is found by subtracting U/2 from the corresponding lower class limit and the upper class boundary is found by adding U/2 to the corresponding upper class limit.
- **Class width**: the difference between the upper and lower class boundaries of any class. It is also the difference between the lower limits of any two consecutive classes or the difference between any two consecutive class marks.

- **Class mark (Mid points):** it is the average of the lower and upper class limits or the average of upper and lower class boundary.
- **Cumulative frequency:** is the number of observations less than/more than or equal to a specific value.
- **Cumulative frequency above:** it is the total frequency of all values greater than or equal to the lower class boundary of a given class.
- **Cumulative frequency blow:** it is the total frequency of all values less than or equal to the upper class boundary of a given class.
- **Cumulative Frequency Distribution (CFD):** it is the tabular arrangement of class interval together with their corresponding cumulative frequencies. It can be more than or less than type, depending on the type of cumulative frequency used.
- **Relative frequency (rf):** it is the frequency divided by the total frequency.
- **Relative cumulative frequency (rcf):** it is the cumulative frequency divided by the total frequency.

Guidelines for classes

1. There should be between 5 and 20 classes.
2. The classes must be mutually exclusive. This means that no data value can fall into two different classes
3. The classes must be all inclusive or exhaustive. This means that all data values must be included.
4. The classes must be continuous. There are no gaps in a frequency distribution.
5. The classes must be equal in width. The exception here is the first or last class. It is possible to have a "below ..." or "... and above" class. This is often used with ages.

Steps for constructing Grouped frequency Distribution

1. Find the largest and smallest values
2. Compute the Range(R) = Maximum - Minimum
3. Select the number of classes desired, usually between 5 and 20 or use Sturges rule $k = 1 + 3.32 \log n$ where *k is number of classes desired and n is total number of observation.*
4. Find the class width by dividing the range by the number of classes and rounding up, not off. $w = \frac{R}{k}$.
5. Pick a suitable starting point less than or equal to the minimum value. The starting point is called the lower limit of the first class. Continue to add the class width to this lower limit to get the rest of the lower limits.
6. To find the upper limit of the first class, subtract U from the lower limit of the second class. Then continue to add the class width to this upper limit to find the rest of the upper limits.
7. Find the boundaries by subtracting U/2 units from the lower limits and adding U/2 units from the upper limits. The boundaries are also half-way between the upper limit of one class and the lower limit of the next class. !may not be necessary to find the boundaries.
8. Tally the data.
9. Find the frequencies.

10. Find the cumulative frequencies. Depending on what you're trying to accomplish, it may not be necessary to find the cumulative frequencies.
11. If necessary, find the relative frequencies and/or relative cumulative frequencies

Example:

Construct a frequency distribution for the following data.

11	29	6	33	14	31	22	27	19	20
18	17	22	38	23	21	26	34	39	27

Solutions:

Step 1: Find the highest and the lowest value H=39, L=6

Step 2: Find the range; R=H-L=39-6=33

Step 3: Select the number of classes desired using Sturges formula;

$$k = 1 + 3.32 \log n = 1 + 3.32 \log (20) = 5.32 = 6 \text{(rounding up)}$$

Step 4: Find the class width; w=R/k=33/6=5.5=6 (rounding up)

Step 5: Select the starting point, let it be the minimum observation.

- 6, 12, 18, 24, 30, 36 are the lower class limits.

Step 6: Find the upper class limit; e.g. the first upper class=12-U=12-1=11

- 11, 17, 23, 29, 35, 41 are the upper class limits.

So combining step 5 and step 6, one can construct the following classes.

Class limits
6 – 11
12 – 17
18 – 23
24 – 29
30 – 35
36 – 41

Step 7: Find the class boundaries;

E.g. for class 1 Lower class boundary=6-U/2=5.5

Upper class boundary =11+U/2=11.5

- Then continue adding w on both boundaries to obtain the rest boundaries. By doing so one can obtain the following classes.

Class boundary
5.5 – 11.5
11.5 – 17.5
17.5 – 23.5
23.5 – 29.5
29.5 – 35.5
35.5 – 41.5

Step 8: tally the data.

Step 9: Write the numeric values for the tallies in the frequency column.

Step 10: Find cumulative frequency.

Step 11: Find relative frequency or/and relative cumulative frequency.

The complete frequency distribution follows:

Class limit	Class boundary	Class Mark	Tally	Freq.	Cf (less than type)	Cf (more than type)	rf.	rcf (less than type
6 – 11	5.5 – 11.5	8.5	//	2	2	20	0.10	0.10
12 – 17	11.5 – 17.5	14.5	//	2	4	18	0.10	0.20
18 – 23	17.5 – 23.5	20.5	/// //	7	11	16	0.35	0.55
24 – 29	23.5 – 29.5	26.5	////	4	15	9	0.20	0.75
30 – 35	29.5 – 35.5	32.5	///	3	18	5	0.15	0.90
36 – 41	35.5 – 41.5	38.5	//	2	20	2	0.10	1.00

Diagrammatic and Graphic presentation of data.

These are techniques for presenting data in visual displays using geometric and pictures.

Importance:

- They have greater attraction.
- They facilitate comparison.
- They are easily understandable.

-Diagrams are appropriate for presenting discrete data.

-The three most commonly used diagrammatic presentation for discrete as well as qualitative data are:

- Pie charts
- Pictogram
- Bar charts

Pie chart

A pie chart is a circle that is divided in to sections or wedges according to the percentage of frequencies in each category of the distribution. The angle of the sector is obtained using:

$$Angle of\ sector = \frac{Value of the part}{the whole quantity} * 360$$

Example: Draw a suitable diagram to represent the following population in a town.

Men	Women	Girls	Boys
2500	2000	4000	1500

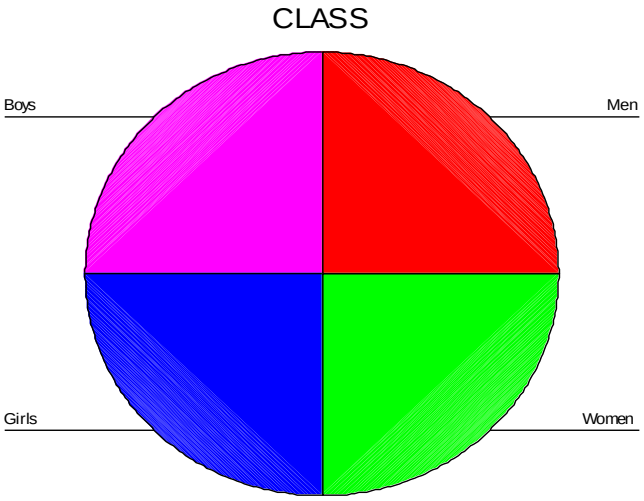
Solutions:

Step 1: Find the percentage.

Step 2: Find the number of degrees for each class.

Step 3: Using a protractor and compass, graph each section and write its name corresponding percentage.

Class	Frequency	Percent	Degree
Men	2500	25	90
Women	2000	20	72
Girls	4000	40	144
Boys	1500	15	54



Pictogram

In these diagram, we represent data by means of some picture symbols. We decide abut a suitable picture to represent a definite number of units in which the variable is measured.

Example: draw a pictogram to represent the following population of a town.

Year	1989	1990	1991	1992
Population	2000	3000	5000	7000

Bar Charts:

- A set of bars (thick lines or narrow rectangles) representing some magnitude over time space.
- They are useful for comparing aggregate over time space.
- Bars can be drawn either vertically or horizontally.
- There are different types of bar charts. The most common being :
    - Simple bar chart
    - Deviation o0r two way bar chart
    - Broken bar chart
    - Component or sub divided bar chart.
    - Multiple bar charts.

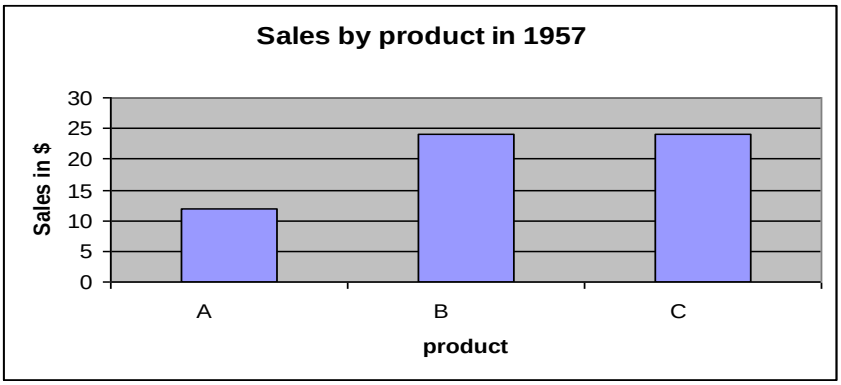
Simple Bar Chart

-Are used to display data on one variable.
-They are thick lines (narrow rectangles) having the same breadth. The magnitude of a quantity is represented by the height /length of the bar.
Example: The following data represent sale by product, 1957- 1959 of a given company for three products A, B, C.

Product	Sales($) In 1957	Sales($) In 1958	Sales($) In 1959
A	12	14	18

B	24	21	18
C	24	35	54

Solutions:



Component Bar chart

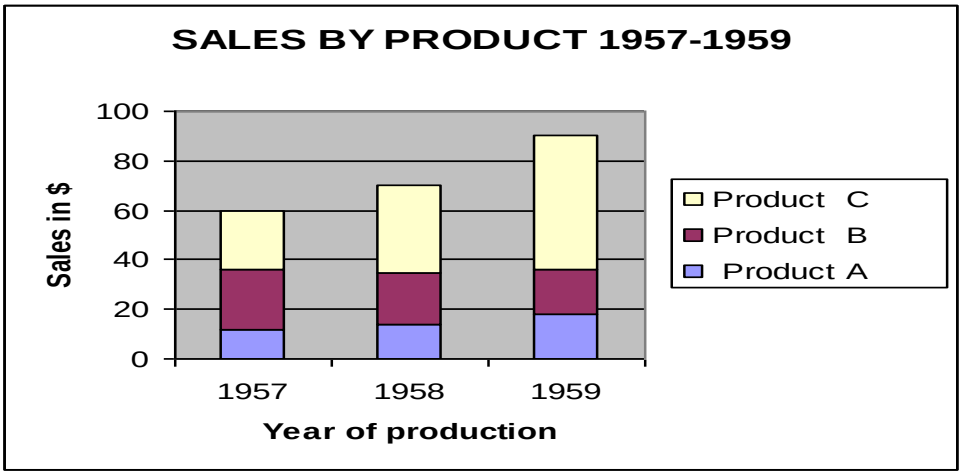
-When there is a desire to show how a total (or aggregate) is divided in to its component parts, we use component bar chart.

-The bars represent total value of a variable with each total broken in to its component parts and different colours or designs are used for identifications

Example:

Draw a component bar chart to represent the sales by product from 1957 to 1959.

Solutions:



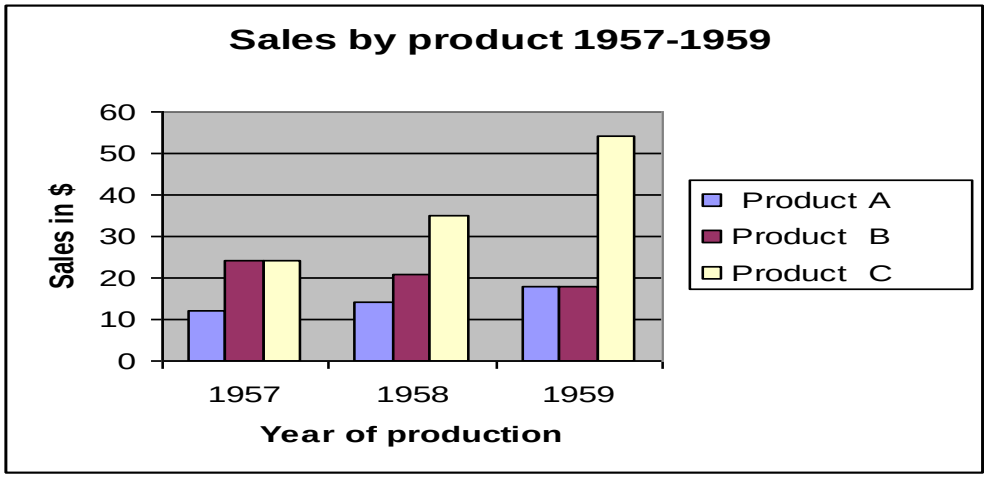
Multiple Bar charts

- These are used to display data on more than one variable.
- They are used for comparing different variables at the same time.

Example:

Draw a component bar chart to represent the sales by product from 1957 to 1959.

Solutions:



Graphical Presentation of data

- The histogram, frequency polygon and cumulative frequency graph or ogive are most commonly applied graphical representation for continuous data.

Procedures for constructing statistical graphs:

- Draw and label the X and Y axes.
- Choose a suitable scale for the frequencies or cumulative frequencies and label it on the Y axes.
- Represent the class boundaries for the histogram or ogive or the mid points for the frequency polygon on the X axes.
- Plot the points.
- Draw the bars or lines to connect the points.

Histogram

A graph which displays the data by using vertical bars of various heights to represent frequencies. Class boundaries are placed along the horizontal axes. Class marks and class limits are some times used as quantity on the X axes.

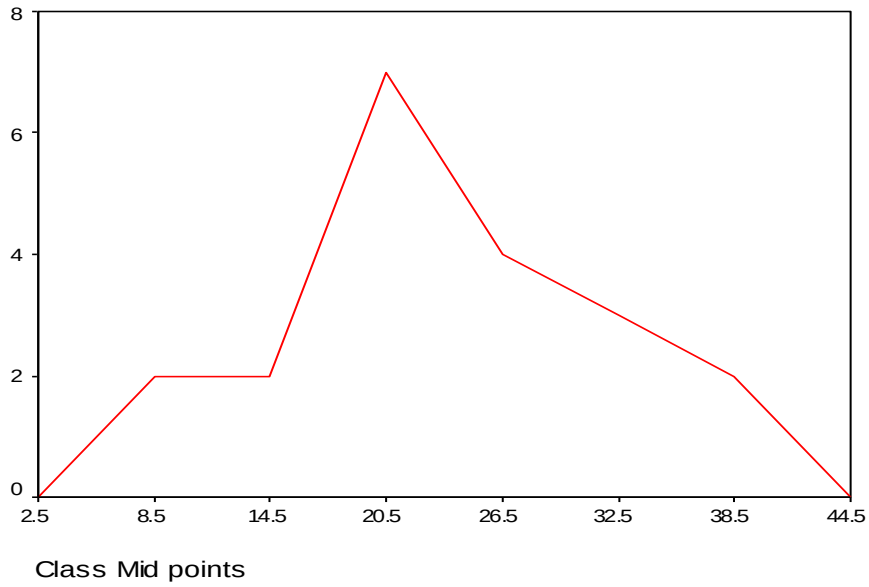
Example: Construct a histogram to represent the previous data (example *).

Frequency Polygon:

A line graph. The frequency is placed along the vertical axis and classes mid points are placed along the horizontal axis. It is customer to the next higher and lower class interval with corresponding frequency of zero, this is to make it a complete polygon.

Example: Draw a frequency polygon for the above data (example *).

Solutions:



Ogive (cumulative frequency polygon)

- A graph showing the cumulative frequency (less than or more than type) plotted against upper or lower class boundaries respectively. That is class boundaries are plotted along the horizontal axis and the corresponding cumulative frequencies are plotted along the vertical axis. The points are joined by a free hand curve.

Example: Draw an ogive curve(less than type) for the above data.

(Example *)

CHAPTER 3

3. MEASURES OF CENTERAL TENDENCY

Introduction

- When we want to make comparison between groups of numbers it is good to have a single value that is considered to be a good representative of each group. This single value is called the **average** of the group. Averages are also called measures of central tendency.
- An average which is representative is called typical average and an average which is not representative and has only a theoretical value is called a descriptive average

Importance:

- To comprehend the data easily.
- To facilitate comparison.
- To make further statistical analysis.

The Summation Notation:

- Let $X_1, X_2, X_3 \ldots X_N$ be a number of measurements where N is the total number of observation and X_i is i^{th} observation.
- Very often in statistics an algebraic expression of the form $X_1+X_2+X_3+...+X_N$ is used in a formula to compute a statistic. It is tedious to write an expression like this very often, so mathematicians have developed a shorthand notation to represent a sum of scores, called the summation notation.
- The symbol $\sum_{i=1}^{N} X_i$ is a mathematical shorthand for $\sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N$

The expression is read, "the sum of X sub i from i equals 1 to N." It means "add up all the numbers."

Example: Suppose the following were scores made on the first homework assignment for five students in the class: 5, 7, 7, 6, and 8. In this example set of five numbers, where N=5, the summation could be written:

$$\sum_{i=1}^{5} X_i = X_1 + X_2 + X_3 + X_4 + X_5 = 5 + 7 + 7 + 6 + 8 = 33$$

The "i=1" in the bottom of the summation notation tells where to begin the sequence of summation. If the expression were written with "i=3", the summation would start with the third number in the set. For example:

$$\sum_{i=3}^{N} X_i = X_3 + X_4 + \ldots + X_N$$

In the example set of numbers, this would give the following result:

$$\sum_{i=3}^{N} X_i = X_3 + X_4 + X_5 = 7 + 6 + 8 = 21$$

The "N" in the upper part of the summation notation tells where to end the sequence of summation. If there were only three scores then the summation and example would be:

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 5 + 7 + 7 = 21$$

Sometimes if the summation notation is used in an expression and the expression must be written a number of times, as in a proof, then a shorthand notation for the shorthand notation is employed. When the summation sign "" is used without additional notation, then "i=1" and "N" are assumed.

For example:

$$\sum X = \sum_{i=1}^{N} X_i = X_1 + X_2 + \ldots + X_N$$

PROPERTIES OF SUMMATION

1. $\sum_{i=1}^{n} k = nk$ where k is any constant
2. $\sum_{i=1}^{n} kX_i = k\sum_{i=1}^{n} X_i$ where k is any constant
3. $\sum_{i=1}^{n} (a + bX_i) = na + b\sum_{i=1}^{n} X_i$ where a and b are any constant
4. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$
5. $$\sum_{i=1}^{N} (X_i * Y_i) = (X_1 * Y_1) + (X_2 * Y_2) + \ldots + (X_N * Y_N)$$

Example: considering the following data determine

X	Y
5	6
7	7
7	8
6	7
8	8

a) $\sum_{i=1}^{5} X_i$

b) $\sum_{i=1}^{5} Y_i$

c) $\sum_{i=1}^{5} 10$

d) $\sum_{i=1}^{5} (X_i + Y_i)$

e) $\sum_{i=1}^{5} (X_i - Y_i)$

f) $\sum_{i=1}^{5} X_i Y_i$

g) $\sum_{i=1}^{5} X_i^{\,2}$

h) $(\sum_{i=1}^{5} X_i)(\sum_{i=1}^{5} Y_i)$

Solutions:

a) $\sum_{i=1}^{5} X_i = 5+7+7+6+8 = 33$

b) $\sum_{i=1}^{5} Y_i = 6+7+8+7+8 = 36$

c) $\sum_{i=1}^{5} 10 = 5*10 = 50$

d) $\sum_{i=1}^{5} (X_i + Y_i) = (5+6)+(7+7)+(7+8)+(6+7)+(8+8) = 69 = 33+36$

e) $\sum_{i=1}^{5} (X_i - Y_i) = (5-6)+(7-7)+(7-8)+(6-7)+(8-8) = -3 = 33-36$

f) $\sum_{i=1}^{5} X_i Y_i = 5*6+7*7+7*8+6*7+8*8 = 241$

g) $\sum_{i=1}^{5} X_i^{\,2} = 5^2+7^2+7^2+6^2+8^2 = 223$

h) $(\sum_{i=1}^{5} X_i)(\sum_{i=1}^{5} Y_i) = 33*36 = 1188$

- **Properties of measures of central tendency (a typical average should posses the following)**
  - It should be rigidly defined.
  - It should be based on all observation under investigation.
  - It should be as little as affected by extreme observations.

- It should be capable of further algebraic treatment.
- It should be as little as affected by fluctuations of sampling.
- It should be ease to calculate and simple to understand.

Types of measures of central tendency

There are several different measures of central tendency; each has its advantage and disadvantage.

- The Mean (Arithmetic, Geometric and Harmonic)
- The Mode
- The Median
- Quintiles (Quartiles, Deciles and Percentiles)

The choice of these averages depends up on which best fit the property under discussion.

The Arithmetic Mean

- Is defined as the sum of the magnitude of the items divided by the number of items.
- The mean of X_1, X_2 ,X_3 …X_n is denoted by A.M ,m or $\overline{X}$ and is given by:

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

$$\Rightarrow \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- If X_1 occurs f_1 times
- If X_2occurs f_2 times
- If X_n occurs f_n times

Then the mean will be $\overline{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}$, where k is the number of classes and

$$\sum_{i=1}^{k} f_i = n$$

Example: Obtain the mean of the following number

2, 7, 8, 2, 7, 3, 7

Solution:

X_i	f_i	$X_i f_i$
2	2	4
3	1	3
7	3	21
8	1	8
Total	7	36

$$\bar{X} = \frac{\sum_{i=1}^{4} f_i X_i}{\sum_{i=1}^{4} f_i} = \frac{36}{7} = 5.15$$

Arithmetic Mean for Grouped Data

If data are given in the shape of a continuous frequency distribution, then the mean is obtained as follows:

$$\bar{X} = \frac{\sum_{i=1}^{k} f_i X_i}{\sum_{i=1}^{k} f_i}, Where$$

X_i =the class mark of the i^{th} class and f_i = the frequency of the i^{th} class

Example: calculate the mean for the following age distribution.

Class	frequency
6- 10	35
11- 15	23
16- 20	15
21- 25	12
26- 30	9
31- 35	6

Solutions:

- First find the class marks
- Find the product of frequency and class marks
- Find mean using the formula.

Class	f_i	X_i	$X_i f_i$
6- 10	35	8	280
11- 15	23	13	299
16- 20	15	18	270
21- 25	12	23	276
26- 30	9	28	252
31- 35	6	33	198
Total	100		1575

$$\bar{X} = \frac{\sum_{i=1}^{6} f_i X_i}{\sum_{i=1}^{6} f_i} = \frac{1575}{100} = 15.75$$

If the values in a series or mid values of a class are large enough, coding of values is a good device to simplify the calculations**.**

Special properties of Arithmetic mean

1. The sum of the deviations of a set of items from their mean is always zero. i.e. $\sum_{i=1}^{n}(X_i - \overline{X}) = 0.$
2. The sum of the squared deviations of a set of items from their mean is the minimum. i.e. $\sum_{i=1}^{n}(Xi - \overline{X})^2 < \sum_{i=1}^{n}(X_i - A)^2, A \neq \overline{X}$
3. If $\overline{X}_1$ is the mean of n_1 observations

   If $\overline{X}_2$ is the mean of n_2 observations

   .

   .

   If $\overline{X}_k$ is the mean of n_k observations

   Then the mean of all the observation in all groups often called the combined mean is given by:

$$\overline{X}_c = \frac{\overline{X}_1 n_1 + \overline{X}_2 n_2 + \ldots + \overline{X}_k n_k}{n_1 + n_2 + \ldots n_k} = \frac{\sum_{i=1}^{k} \overline{X}_i n_i}{\sum_{i=1}^{k} n_i}$$

Example: In a class there are 30 females and 70 males. If females averaged 60 in an examination and boys averaged 72, find the mean for the entire class.

Solutions:

Females	*Males*
$\overline{X}_1 = 60$	$\overline{X}_2 = 72$
$n_1 = 30$	$n_2 = 70$

$$\overline{X}_c = \frac{\overline{X}_1 n_1 + \overline{X}_2 n_2}{n_1 + n_2} = \frac{\sum_{i=1}^{2} \overline{X}_i n_i}{\sum_{i=1}^{2} n_i}$$

$$\Rightarrow \overline{X}_c = \frac{30(60) + 70(72)}{30 + 70} = \frac{6840}{100} = 68.40$$

4. If a wrong figure has been used when calculating the mean the correct mean can be obtained without repeating the whole process using:

$$\mathbf{CorrectMean = WrongMean + \frac{(CorrectValue - Wrong Va)}{n}}$$

Where n is total number of observations.

Example: An average weight of 10 students was calculated to be 65. Latter it was discovered that one weight was misread as 40 instead of 80 kg. Calculate the correct average weight.

Solutions:

$$\mathbf{CorrectMean = WrongMean + \frac{(CorrectValue - WrongValue)}{n}}$$

$$\mathbf{CorrectMean = 65 + \frac{(80-40)}{10} = 65 + 4 = 69 k.g.}$$

5. The effect of transforming original series on the mean.
   a) If a constant *k* is added/ subtracted to/from every observation then the new mean will be *the old mean*± *k* respectively.
   b) If every observations are multiplied by a constant *k* then the new mean will be *k*old mean*

Example:

1. The mean of n Tetracycline Capsules X_1, X_2, …,X_n are known to be 12 gm. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ (i = 1, 2, …, n) then what will be the mean of the new set of capsules

Solutions:

$$\mathbf{NewMean = 2 * OldMean - 0.5 = 2 * 12 - 0.5 = 23.5}$$

2. The mean of a set of numbers is 500.
   a) If 10 is added to each of the numbers in the set, then what will be the mean of the new set?
   b) If each of the numbers in the set are multiplied by -5, then what will be the mean of the new set?

Solutions:

a).$\mathbf{NewMean = OldMean + 10 = 500 + 10 = 510}$

b).$\mathbf{New\ Mean = 5 * Old\ Mean = 5 * 500 = -2500}$

Weighted Mean

- When a proper importance is desired to be given to different data a weighted mean is appropriate.
- Weights are assigned to each item in proportion to its relative importance.
- Let X_1, X_2, …X_n be the value of items of a series and W_1, W_2, …W_n their corresponding weights , then the weighted mean denoted $\overline{X}_w$ is defined as:

$$\overline{\mathbf{X}}_{\mathbf{w}} = \frac{\sum_{i=1}^{n} \mathbf{X_i W_i}}{\sum_{i-1}^{n} \mathbf{W_i}}$$

Example:

A student obtained the following percentage in an examination:
English 60, Biology 75, Mathematics 63, Physics 59, and chemistry 55.Find the students weighted arithmetic mean if weights 1, 2, 1, 3, 3 respectively are allotted to the subjects.

Solutions:

$$\overline{X}_w = \frac{\sum_{i=1}^{5} X_i W_i}{\sum_{i-1}^{5} W_i} = \frac{60*1+75*2+63*1+59*3+55*3}{1+2+1+3+3} = \frac{615}{10} = 61.5$$

Merits and Demerits of Arithmetic Mean

Merits:

- It is rigidly defined.
- It is based on all observation.
- It is suitable for further mathematical treatment.
- It is stable average, i.e. it is not affected by fluctuations of sampling to some extent.
- It is easy to calculate and simple to understand.

Demerits:

- It is affected by extreme observations.
- It cannot be used in the case of open end classes.
- It cannot be determined by the method of inspection.
- It cannot be used when dealing with qualitative characteristics, such as intelligence, honesty, beauty.
- It can be a number which does not exist in a serious.
- Sometimes it leads to wrong conclusion if the details of the data from which it is obtained are not available.
- It gives high weight to high extreme values and less weight to low extreme values.

The Geometric Mean

- ☞ The geometric mean of a set of n observation is the n^{th} root of their product.
- ☞ The geometric mean of X_1, X_2 ,X_3 …X_n is denoted by G.M and given by:

$$G.M = \sqrt[n]{X_1 * X_2 * \ldots * X_n}$$

- ☞ Taking the logarithms of both sides

$$\log(G.M) = \log(\sqrt[n]{X_1 * X_2 * \ldots * X_n}) = \log(X_1 * X_2 * \ldots * X_n)^{\frac{1}{n}}$$

$$\Rightarrow \log(G.M) = \frac{1}{n}\log(X_1 * X_2 * \ldots * X_n) = \frac{1}{n}(\log X_1 + \log X_2 + \ldots + \log X_n)$$

$$\Rightarrow \log(G.M) = \frac{1}{n}\sum_{i=1}^{n}\log X_i$$

$\Rightarrow$ **The logarithm of the G.M of a set of observation is the arithmetic mean of their logarithm.**

$$\Rightarrow G.M = Anti\log\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right)$$

Example:

Find the G.M of the numbers 2, 4, 8.

Solutions:

$$G.M = \sqrt[n]{X_1 * X_2 * \ldots * X_n} = \sqrt[3]{2*4*8} = \sqrt[3]{64} = 4$$

Remark: The Geometric Mean is useful and appropriate for finding averages of ratios.

The Harmonic Mean

The harmonic mean of $X_1, X_2, X_3 \ldots X_n$ is denoted by H.M and given by:

$$H.M = \frac{n}{\sum_{i=1}^{n} \frac{1}{X_i}}$$, This is called simple harmonic mean.

In a case of frequency distribution:

$$H.M = \frac{n}{\sum_{i=1}^{k} \frac{f_i}{X_i}} \quad , \quad n = \sum_{i=1}^{k} f_i$$

If observations X_1, $X_2 \ldots$ Xn have weights W_1, W_2 …Wn respectively, then their harmonic mean is given by

$$H.M = \frac{\sum_{i=1}^{n} W_i}{\sum_{i=1}^{n} W_i / X_i}$$, This is called Weighted Harmonic Mean.

Remark: The Harmonic Mean is useful and appropriate in finding average speeds and average rates.

Example: A cyclist pedals from his house to his college at speed of 10 km/hr and back from the college to his house at 15 km/hr. Find the average speed.

Solution: Here the distance is constant

➔The simple H.M is appropriate for this problem.

X_1= 10km/hr X_2=15km/hr

$$H.M = \frac{2}{\frac{1}{10} + \frac{1}{15}} = 12 km/hr$$

The Mode

- Mode is a value which occurs most frequently in a set of values
- The mode may not exist and even if it does exist, it may not be unique.
- In case of discrete distribution the value having the maximum frequency is the model value.

Examples:

1. Find the mode of 5, 3, 5, 8, 9
   Mode =5
2. Find the mode of 8, 9, 9, 7, 8, 2, and 5.
   It is a bimodal Data: 8 and 9

3. Find the mode of 4, 12, 3, 6, and 7.
   No mode for this data.

- The mode of a set of numbers $X_1, X_2, \ldots, X_n$ is usually denoted by $\hat{X}$.

Mode for Grouped data

If data are given in the shape of continuous frequency distribution, the mode is defined as:

$$\hat{\mathbf{X}} = \mathbf{L}_{mo} + \mathbf{w}\left(\frac{\Delta_1}{\Delta_1 + \Delta_2}\right)$$

Where:

$\hat{X} = the \bmod e\ of\ the\ distribution$

$w = the\ size\ of\ the \bmod al\ class$

$\Delta_1 = f_{mo} - f_1$

$\Delta_2 = f_{mo} - f_2$

$f_{mo} = frequency\ of\ the \bmod al\ class$

$f_1 = frequency\ of\ the\ class\ preceeding\ the \bmod al\ class$

$f_2 = frequency\ of\ the\ class\ \ following\ the \bmod al\ class$

Note: The modal class is a class with the highest frequency.

Example: Following is the distribution of the size of certain farms selected at random from a district. Calculate the mode of the distribution.

Size of farms	No. of farms
5-15	8
15-25	12
25-35	17
35-45	29
45-55	31
55-65	5
65-75	3

Solutions:

$\mathbf{45-55\ is\ the} \bmod \mathbf{al\ class}, \sin \mathbf{ce\ it\ is\ a\ class\ with\ the\ highest\ frequency}.$

$\mathbf{L_{mo} = 45}$

$\mathbf{w = 10}$

$\mathbf{\Delta_1 = f_{mo} - f_1 = 2}$

$\mathbf{\Delta_2 = f_{mo} - f_2 = 26}$

$\mathbf{f_{mo} = 31}$

$\mathbf{f_1 = 29}$

$\mathbf{f_2 = 5}$

$$\Rightarrow \hat{X} = 45 + 10\left(\frac{2}{2+26}\right)$$

$$= 45.71$$

Merits and Demerits of Mode

Merits:

- It is not affected by extreme observations.
- Easy to calculate and simple to understand.
- It can be calculated for distribution with open end class

Demerits:

- It is not rigidly defined.
- It is not based on all observations
- It is not suitable for further mathematical treatment.
- It is not stable average, i.e. it is affected by fluctuations of sampling to some extent.
- Often its value is not unique.

Note: being the point of maximum density, mode is especially useful in finding the most popular size in studies relating to marketing, trade, business, and industry. It is the appropriate average to be used to find the ideal size.

The Median

- In a distribution, median is the value of the variable which divides it in to two equal halves.
- In an ordered series of data median is an observation lying exactly in the middle of the series. It is the middle most value in the sense that the number of values less than the median is equal to the number of values greater than it.

-If $X_1, X_2 \ldots Xn$ be the observations, then the numbers arranged in ascending order will be $X_{[1]}, X_{[2]} \ldots X_{[n]}$, where $X_{[i]}$ is i^{th} smallest value.

$\Rightarrow X_{[1]} < X_{[2]} < \ldots < X_{[n]}$

-Median is denoted by $\tilde{X}$.

Median for ungrouped data

$$\tilde{X} = \begin{cases} X_{[(n+1)/2]} & ,\text{If } n \text{ is odd.} \\ \frac{1}{2}(X_{[n/2]} + X_{[(n/2)+1]}), & \text{If } n \text{ is even} \end{cases}$$

Example: Find the median of the following numbers.

a) 6, 5, 2, 8, 9, 4.
b) 2, 1, 3, 5, 8.

Solutions:

a) First order the data: 2, 4, 5, 6, 8, 9
Here n=6

$$\tilde{X} = \frac{1}{2}(X_{[\frac{n}{2}]} + X_{[\frac{n}{2}+1]})$$

$$= \frac{1}{2}(X_{[3]} + X_{[4]})$$

$$= \frac{1}{2}(5+6) = 5.5$$

b) Order the data :1, 2, 3, 5, 8

Here n=5

$$\tilde{X} = X_{[\frac{n+1}{2}]}$$

$$= X_{[3]}$$

$$= 3$$

Median for grouped data If data are given in the shape of continuous frequency distribution, the median is defined as:

$$\tilde{X} = L_{med} + \frac{w}{f_{med}}(\frac{n}{2} - c)$$

Where:

L_{med} = lower class boundary of the median class.

w = the size of the median class

n = total number of observations.

c = the cumulative frequency (less than type) preceeding the median class.

f_{med} = the frequency of the median class.

Remark:

The median class is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{n}{2}$.

Example: Find the median of the following distribution.

Class	Frequency
40-44	7
45-49	10
50-54	22
55-59	15
60-64	12
65-69	6
70-74	3

Solutions:

- First find the less than cumulative frequency.
- Identify the median class.
- Find median using formula.

Class	Frequency	Cumu.Freq(less than type)
40-44	7	7
45-49	10	17
50-54	22	39
55-59	15	54
60-64	12	66
65-69	6	72
70-74	3	75

$$\frac{n}{2}=\frac{75}{2}=37.5$$

$$39 \text{ is the first cumulative frequency to be greater than or equal to } 37.5$$

$$\Rightarrow 50-54 \text{ is the median class.}$$

$$L_{med}=49.5, \quad w=5$$
$$n=75, \quad c=17, \quad f_{med}=22$$

$$\Rightarrow \tilde{X}=L_{med}+\frac{w}{f_{med}}\left(\frac{n}{2}-c\right)$$
$$=49.5+\frac{5}{22}(37.5-17)$$
$$=54.16$$

Merits and Demerits of Median

Merits:

- Median is a positional average and hence not influenced by extreme observations.
- Can be calculated in the case of open end intervals.
- Median can be located even if the data are incomplete.

Demerits:

- It is not a good representative of data if the number of items is small.
- It is not amenable to further algebraic treatment.
- It is susceptible to sampling fluctuations.

Quantiles

When a distribution is arranged in order of magnitude of items, the median is the value of the middle term. Their measures that depend up on their positions in distribution quartiles, deciles, and percentiles are collectively called quantiles.

Quartiles:

- Quartiles are measures that divide the frequency distribution in to four equal parts.
- The value of the variables corresponding to these divisions are denoted Q_1, Q_2, and Q_3 often called the first, the second and the third quartile respectively.
- Q_1 is a value which has 25% items which are less than or equal to it. Similarly Q_2 has 50%items with value less than or equal to it and Q_3 has 75% items whose values are less than or equal to it.

Calculating quartiles for raw data

- To calculate the three quartiles from the raw data, we must arranged the data from least to highest 1st if the data are arranged in increasing order ,then

$$Q_i = \tfrac{i}{4}(n+1)^{th} value, i = 1,2,3, then$$

$$Q_1 = \tfrac{1}{4}(n+1)^{th} vlaue$$

$$Q_2 = \tfrac{2}{4}(n+1)^{th} vlaue$$

$$Q_3 = \tfrac{3}{4}(n+1)^{th} v\ l\ a\ \iota$$, where n is number of observations.

E.g. the following data shows the age of 30 sampled patients in JUSH

6,9,11,14,16,17,18,21,22,22,22,22,23,25,25,26,27,28,28,32,33,34,34,36,39,39,41,45,46,49 find the lower middle and upper quartiles for the above data.

Solution:

1st order the data (if it hasn't been ordered)

6,9,11,14,16,17,18,21,22,22,22,22,23,25,25,26,27,28,28,32,33,34,34,36,39,39,41,45,46,49

n =30, $Q_1 = \tfrac{1}{4}(n+1)^{th} value = \tfrac{1}{4}(30+1)^{th} value =$ 7.75th value =7th value +0.75(8th value - 7th value)

18+0.75(21-18) = 18+2.25 =20.25

This implies one fourth of the patients (25%) age are below 20.5 years.

$Q_1 = \quad \tfrac{2}{4}(n+1)^{th} value =$

$\tfrac{2}{4}(30+1)^{th} value = 2(7.75)^{th} value = 15.5^{th} value$

$= 15^{th} value + 0.5(16^{th} value - 15^{th} value) = 25 + 0.5(26-25) = 25.5$

This implies that half (50%) of the patients age is below 25.5 years.

Calculating quartiles for grouped data: To find Q_i (i=1, 2, 3) from grouped frequency distribution, we count $\frac{in}{4}$ of the classes beginning from the lowest class.

So to calculate the quartiles, We have the following formula,

$$Q_i = L_{Q_i} + \frac{w}{f_{Q_i}}(\frac{in}{4} - c) \quad , i = 1,2,3$$

Where:

$L_{Q_i} = lower\ class\ boundary\ of\ the\ quartile\ class.$

$w = the\ size\ of\ the\ quartile\ class$

$n = \ total\ number\ of\ observations.$

$c = the\ cumulative\ frequency\ (less\ than\ type)\ preceeding\ the\ quartile\ class.$

$f_{Q_i} = the\ frequency\ of\ the\ quartile\ class.$

Remark: The quartile class (class containing Q_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{in}{4}$.

Deciles:

- Deciles are measures that divide the frequency distribution in to ten equal parts.
- The values of the variables corresponding to these divisions are denoted D_1, D_2,.. , D9 often called the first, the second,…, the ninth deciles respectively.

Calculating Deciles for raw data

- To calculate the nine deciles from the raw data, we must arranged the data from least to highest 1st if the data are not arranged in increasing order ,then

$D_i = \tfrac{i}{10}(n+1)^{th} value, i = 1,2,3,....9 then$

$D_1 = \tfrac{1}{10}(n+1)^{th} vlaue$

$D_2 = \tfrac{2}{10}(n+1)^{th} vlaue$

$$D_3 = \frac{3}{10}(n+1)^{th}\,value$$

.
.

$$D_9 = \frac{9}{10}(n+1)^{th}\,value$$, where n is number of observations.

Calculating Deciles For grouped data:

- To find D_i (i=1, 2,..,9) We count $\frac{in}{10}$ of the classes beginning from the lowest class. we have the following formula

$$D_i = L_{D_i} + \frac{w}{f_{D_i}}\left(\frac{iN}{10} - c\right) \quad , i = 1,2,\ldots 9 ,$$

Where:

$L_{D_i} = lower\ class\ boundary\ of\ the\ decile\ class.$

$w = the\ size\ of\ the\ decile\ class$

$n = total\ number\ of\ observations.$

$c = the\ cumulative\ frequency\ (less\ than\ type)\ preceeding\ the\ decile\ class.$

$f_{D_i} = the\ frequency\ of\ the\ decile\ class.$

Remark:

The decile class (class containing D_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{in}{10}$.

Percentiles:

- Percentiles are measures that divide the frequency distribution in to hundred equal parts.
- The values of the variables corresponding to these divisions are denoted P_1, P_2,.. P_{99} often called the first, the second,…,the ninety-ninth percentile respectively.
- To calculate the ninety-nine percentile from the raw data, we must have arranged the data from least to highest 1st if the data are not arranged in increasing order, then

$$p_i = \frac{i}{100}(n+1)^{th} value, i = 1,2,3,....9 then$$

$$p_1 = \frac{1}{100}(n+1)^{th} vlaue$$

$$p_2 = \frac{2}{100}(n+1)^{th} vlaue$$

$$p_3 = {}^{3}\!/_{100}(n+1)^{th}\,value$$

.

.

$$p_{99} = {}^{99}\!/_{100}(n+1)^{th}\,value$$, where n is number of observations.

Calculating For grouped data: To find P_i (i=1, 2... 99) we count $\frac{in}{100}$ of the classes beginning from the lowest class. We have the following formula

$$P_i = L_{P_i} + \frac{w}{f_{P_i}}\left(\frac{in}{100} - c\right) \quad , i = 1,2,...,99$$

Where:

$L_{P_i} = lower\ class\ boundary\ of\ the\ percentile\ class.$

$w = the\ size\ of\ the\ percentile\ class$

$N = total\ number\ of\ observations.$

$c = the\ cumulative\ frequency\ (less\ than\ type)\ preceeding\ the\ percentile\ class.$

$f_{P_i} = the\ frequency\ of\ the\ percentile\ class.$

Remark: The percentile class (class containing P_i) is the class with the smallest cumulative frequency (less than type) greater than or equal to $\frac{in}{100}$.

Example: Considering the following distribution

Calculate:

a) All quartiles.
b) The 7^{th} decile.
c) The 90^{th} percentile.

Values	Frequency
140- 150	17
150- 160	29
160- 170	42
170- 180	72
180- 190	84
190- 200	107
200- 210	49
210- 220	34
220- 230	31
230- 240	16
240- 250	12

Solutions:

- First find the less than cumulative frequency.
- Use the formula to calculate the required quantile.

Values	Frequency	Cum.Freq(less

		than type)
140- 150	17	17
150- 160	29	46
160- 170	42	88
170- 180	72	160
180- 190	84	244
190- 200	107	351
200- 210	49	400
210- 220	34	434
220- 230	31	465
230- 240	16	481
240- 250	12	493

a. Quartiles:

i. Q_1

- determine the class containing the first quartile.

$$\frac{n}{4} = 123.25$$

$$\Rightarrow 170-180 \ \textit{is the class containing the first quartile.}$$

$L_{Q_1} = 170$, $w = 10$

$n = 493$, $c = 88$, $f_{Q_1} = 72$

$$\Rightarrow Q_1 = L_{Q_1} + \frac{w}{f_{Q_1}}\left(\frac{n}{4} - c\right)$$

$$= 170 + \frac{10}{72}(123.25 - 88)$$

$$= \underline{\underline{174.90}}$$

ii. Q_2

- determine the class containing the second quartile.

$$\frac{2*n}{4} = 246.5$$

$$\Rightarrow 190-200 \ \textit{is the class containing the second quart.}$$

$L_{Q_2} = 190$, $w = 10$

$n = 493$, $c = 244$, $f_{Q_2} = 107$

$$\Rightarrow Q_2 = L_{Q_2} + \frac{w}{f_{Q_2}}\left(\frac{2*n}{4} - c\right)$$

$$= 170 + \frac{10}{72}(246.5 - 244)$$

$$= \underline{\underline{190.23}}$$

iii. Q_3

- determine the class containing the third quartile.

$$\frac{3*n}{4} = 369.75$$

$$\Rightarrow 200-210 \text{ is the class containing the third quartile.}$$

$$L_{Q_3} = 200, \quad w = 10$$
$$n = 493, \quad c = 351, \quad f_{Q_3} = 49$$

$$\Rightarrow Q_3 = L_{Q_3} + \frac{w}{f_{Q_3}}\left(\frac{3*n}{4} - c\right)$$
$$= 200 + \frac{10}{49}(369.75 - 351)$$
$$= \underline{\underline{203.83}}$$

b. D_7

- determine the class containing the 7th decile.

$$\frac{7*n}{10} = 345.1$$

$$\Rightarrow 190-200 \text{ is the class containing the seventh decile.}$$

$$L_{D_7} = 190, \quad w = 10$$
$$n = 493, \quad c = 244, \quad f_{D_7} = 107$$

$$\Rightarrow D_7 = L_{D_7} + \frac{w}{f_{D_7}}\left(\frac{7*n}{10} - c\right)$$
$$= 190 + \frac{10}{107}(345.1 - 244)$$
$$= \underline{\underline{199.45}}$$

c. P_{90}, determine the class containing the 90th percentile.

$$\frac{90*n}{100} = 443.7$$

$$\Rightarrow 220-230 \text{ is the class containing the } 90^{th} \text{ percentile.}$$

$$L_{P_{90}} = 220, \quad w = 10$$
$$n = 493, \quad c = 434, \quad f_{P_{90}} = 31$$

$$\Rightarrow P_{90} = L_{P_{90}} + \frac{w}{f_{P_{90}}}\left(\frac{90 * n}{100} - c\right)$$

$$= 220 + \frac{10}{31}(443.7 - 433)$$

$$= \underline{\underline{223.13}}$$

CHAPTER 4

4. Measures of Dispersion (Variation)

4.1 Introduction and objectives of measuring Variation

-The scatter or spread of items of a distribution is known as dispersion or variation. In other words the degree to which numerical data tend to spread about an average value is called dispersion or variation of the data.
-Measures of dispersions are statistical measures which provide ways of measuring the extent in which data are dispersed or spread out.
Objectives of measuring Variation:

- To judge the reliability of measures of central tendency
- To control variability itself.
- To compare two or more groups of numbers in terms of their variability.
- To make further statistical analysis.

Absolute and Relative Measures of Dispersion

The measures of dispersion which are expressed in terms of the original unit of a series are termed as *absolute measures*. Such measures are not suitable for comparing the variability of two distributions which are expressed in different *units of measurement* and different average size. Relative measures of dispersions are a ratio or percentage of a measure of absolute dispersion to an appropriate measure of central tendency and are thus pure numbers independent of the *units of measurement*. For comparing the variability of two distributions (even if they are measured in the same unit), we compute the relative measure of dispersion instead of absolute measures of dispersion.

4.2 Types of Measures of Dispersion

Various measures of dispersions are in use. The most commonly used measures of dispersions are:

1) Range and relative range
2) Quartile deviation and coefficient of Quartile deviation
3) Mean deviation and coefficient of Mean deviation
4) Standard deviation ,coefficient of variation and standard scores

The Range (R)

The range is the largest score minus the smallest score. It is a quick and dirty measure of variability, although when a test is given back to students they very often wish to know the range of scores. Because the range is greatly affected by extreme scores, it may give a distorted picture of the scores. The following two distributions have the same range, 13, yet appear to differ greatly in the amount of variability.

Distribution 1:	32	35	36	36	37	38	40	42	42	43	43	45
Distribution 2:	32	32	33	33	33	34	34	34	34	34	35	45

For this reason, among others, the range is not the most important measure of variability.

$$R = L - S \qquad L = \text{largest observant}$$
$$S = \text{smallest observant}$$

Range for grouped data:

If data are given in the shape of continuous frequency distribution, the range is computed as:

$$R = UCL_k - UCL_1, \qquad UCL_k \text{ is upper class limit of the last class.}$$
$$UCL_1 \text{ is lower class limit of the first class.}$$

This is some times expressed as:

$$R = X_k - X_1, \qquad X_k \text{ is class mark of the last class.}$$
$$X_1 \text{ is class mark of the first class}$$

Merits and Demerits of range

Merits:

- It is rigidly defined.
- It is easy to calculate and simple to understand.

Demerits:

- It is not based on all observation.
- It is highly affected by extreme observations.
- It is affected by fluctuation in sampling.
- It is not liable to further algebraic treatment.
- It cannot be computed in the case of open end distribution.
- It is very sensitive to the size of the sample.

Relative Range (RR)

-it is also sometimes called coefficient of range and given by:

$$RR = \frac{L - S}{L + S} = \frac{R}{L + S}$$

Example:

1. Find the relative range of the above two distribution.(exercise!)

2. If the range and relative range of a series are 4 and 0.25 respectively. Then what is the value of:
   a) Smallest observation
   b) Largest observation

Solutions :(2)

$R = 4 \Rightarrow L - S = 4$ _____(1)

$RR = 0.25 \Rightarrow L + S = 16$ _____(2)

Solving (1) *and* (2) *at the same time, one can obtain the following value*

$L = 10$ *and* $S = 6$

The Quartile Deviation (Semi-inter quartile range), Q.D

The inter quartile range is the difference between the third and the first quartiles of a set of items and semi-inter quartile range is half of the inter quartile range.

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

Coefficient of Quartile Deviation (C.Q.D)

$$C.Q.D = \frac{(Q_3 - Q_1/2}{(Q_3 + Q_1)/2} = \frac{2 * Q.D}{Q_3 + Q_1} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

- It gives the average amount by which the two quartiles differ from the median.

Example: Compute Q.D and its coefficient for the following distribution.

Values	Frequency
140- 150	17
150- 160	29
160- 170	42
170- 180	72
180- 190	84
190- 200	107
200- 210	49
210- 220	34
220- 230	31
230- 240	16
240- 250	12

Solutions:

In the previous chapter we have obtained the values of all quartiles as:
Q_1= 174.90, Q_2= 190.23, Q_3=203.83

$$\Rightarrow \quad Q.D = \frac{Q_3 - Q_1}{2} = \frac{203.83 - 174.90}{2} = 14.47$$

$$C.Q.D = \frac{2 * Q.D}{Q_3 + Q_1} = \frac{2 * 14.47}{203.83 + 174.90} = 0.076$$

Remark: Q.D or C.Q.D includes only the middle 50% of the observation.

The Mean Deviation (M.D):

The mean deviation of a set of items is defined as the arithmetic mean of the values of the absolute deviations from a given average. Depending up on the type of averages used we have different mean deviations.

a) Mean Deviation about the mean
   - Denoted by M.D($\bar{X}$) and given by

$$M.D(\bar{X}) = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

   - For the case of frequency distribution it is given as:

$$M.D(\bar{X}) = \frac{\sum_{i=1}^{k} f_i |X_i - \bar{X}|}{n}$$

Steps to calculate M.D ($\bar{X}$):

1. Find the arithmetic mean, $\bar{X}$
2. Find the deviations of each reading from $\bar{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

b) Mean Deviation about the median.
   - Denoted by M.D($\tilde{X}$) and given by

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^{n} |X_i - \tilde{X}|}{n}$$

   - For the case of frequency distribution it is given as:

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^{k} f_i \left| X_i - \tilde{X} \right|}{n}$$

Steps to calculate M.D ($\tilde{X}$):

1. Find the median, $\tilde{X}$
2. Find the deviations of each reading from $\tilde{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

c) Mean Deviation about the mode.

- Denoted by M.D($\hat{X}$) and given by

$$\mathbf{M.D}(\hat{\mathbf{X}}) = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} \left| \mathbf{X_i} - \hat{\mathbf{X}} \right|}{\mathbf{n}}$$

- For the case of frequency distribution it is given as:

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{k} f_i \left| X_i - \hat{X} \right|}{n}$$

Steps to calculate M.D ($\hat{X}$):

1. Find the mode, $\hat{X}$
2. Find the deviations of each reading from $\hat{X}$.
3. Find the arithmetic mean of the deviations, ignoring sign.

Examples:

1. The following are the number of visit made by ten mothers to the local doctor's surgery.
   8, 6, 5, 5, 7, 4, 5, 9, 7, 4
   Find mean deviation about mean, median and mode.
   Solutions:
   First calculate the three averages
   $\bar{X} = 6,\ \tilde{X} = 5.5,\quad \hat{X} = 5$

   Then take the deviations of each observation from these averages.

X_i	4	4	5	5	5	6	7	7	8	9	total
$\left\|X_i - 6\right\|$	2	2	1	1	1	0	1	1	2	3	14

$\lvert X_i - 5.5\rvert$	1.5	1.5	0.5	0.5	0.5	0.5	1.5	1.5	2.5	3.5	14
$\lvert X_i - 5\rvert$	1	1	0	0	0	1	2	2	3	4	14

$$\Rightarrow M.D(\bar{X}) = \frac{\sum_{i=1}^{10} |X_i - 6)|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\tilde{X}) = \frac{\sum_{i=1}^{10} |X_i - 5.5|}{10} = \frac{14}{10} = 1.4$$

$$M.D(\hat{X}) = \frac{\sum_{i=1}^{10} |X_i - 5)|}{10} = \frac{14}{10} = 1.4$$

2. Find mean deviation about mean, median and mode for the following distributions.(exercise)

Class	Frequency
40-44	**7**
45-49	**10**
50-54	**22**
55-59	**15**
60-64	**12**
65-69	**6**
70-74	**3**

Remark: Mean deviation is always minimum about the median.

Coefficient of Mean Deviation (C.M.D)

$$C.M.D = \frac{M.D}{Average \ \ about\, which\, deviations\, are\, taken}$$

$$\Rightarrow C.M.D(\bar{X}) = \frac{M.D(\bar{X})}{\bar{X}}$$

$$C.M.D(\tilde{X}) = \frac{M.D(\tilde{X})}{\tilde{X}}$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}}$$

Example: calculate the C.M.D about the mean, median and mode for the data in example 1 above.

Solutions:

$$C.M.D = \frac{M.D}{Average\ \ about\ which\ deviations\ are\ taken}$$

$$\Rightarrow C.M.D(\bar{X}) = \frac{M.D(\bar{X})}{\bar{X}} = \frac{1.4}{6} = 0.233$$

$$C.M.D(\tilde{X}) = \frac{M.D(\tilde{X})}{\tilde{X}} = \frac{1.4}{5.5} = 0.255$$

$$C.M.D(\hat{X}) = \frac{M.D(\hat{X})}{\hat{X}} = \frac{1.4}{5} = 0.28$$

Exercise: Identify the merits and demerits of Mean Deviation

The Variance

Population Variance

If we divide the variation by the number of values in the population, we get something called the population variance. This variance is the "average squared deviation from the mean".

$$Population\ Varince = \sigma^2 = \frac{1}{N}\sum (X_i - \mu)^2, \quad i = 1,2,.....N$$

For the case of frequency distribution it is expressed as:

$$Population\ Varince = \sigma^2 = \frac{1}{N}\sum f_i (X_i - \mu)^2, \quad i = 1,2,....k$$

Sample Variance

One would expect the sample variance to simply be the population variance with the population mean replaced by the sample mean. However, one of the major uses of statistics is to estimate the corresponding parameter. This formula has the problem that the estimated value isn't the same as the parameter. To counteract this, the sum of the squares of the deviations is divided by one less than the sample size.

$$Sample\,Varince = S^2 = \frac{1}{n-1}\sum (X_i - \overline{X})^2, \quad i = 1,2,.....,n$$

For the case of frequency distribution it is expressed as:

$$Sample\,Varince = S^2 = \frac{1}{n-1}\sum f_i (X_i - \overline{X})^2, \quad i = 1,2,....k$$

We usually use the following short cut formula.

$$S^2 = \frac{\sum_{i=1}^{n} X_i{}^2 - n\overline{X}^2}{n-1}, \quad for\ raw\ data$$

$$S^2 = \frac{\sum_{i=1}^{k} f_i X_i{}^2 - n\overline{X}^2}{n-1}, \quad for\ frequency\ distribution.$$

Standard Deviation

There is a problem with variances. Recall that the deviations were squared. That means that the units were also squared. To get the units back the same as the original data values, the square root must be taken.

$$Populations\tan dard\ deviation = \sigma = \sqrt{\sigma^2}$$

$$Samples\tan dard\ deviation = s = \sqrt{S^2}$$

The following steps are used to calculate the sample standard deviation

1. Find the arithmetic mean.
2. Find the difference between each observation and the mean.
3. Square these differences.
4. Sum the squared differences.

5. Since the data is a sample, divide the number (from step 4 above) by the number of observations minus one, i.e., n-1 (where n is equal to the number of observations in the data set).

6. Square root the result obtained from step 5

Examples: Find the variance and standard deviation of the following sample data

1. 5, 17, 12, 10.
2. The data is given in the form of frequency distribution.

Class	Frequency
40-44	**7**
45-49	**10**
50-54	**22**
55-59	**15**
60-64	**12**
65-69	**6**
70-74	**3**

Solutions:

1. $\overline{X} = 11$

X_i	5	10	12	17	Total
$(X_i - \overline{X})^2$	36	1	1	36	74

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{74}{3} = 24.67.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{24.67} = 4.97.$$

2. $\overline{X} = 55$

X_i(C.M)	42	47	52	57	62	67	72	Total
$f_i(X_i - \overline{X})^2$	1183	640	198	60	588	864	867	4400

$$\Rightarrow S^2 = \frac{\sum_{i=1}^{n} f_i(X_i - \overline{X})^2}{n-1} = \frac{4400}{74} = 59.46.$$

$$\Rightarrow S = \sqrt{S^2} = \sqrt{59.46} = 7.71.$$

Special properties of Standard deviations

1. $\sqrt{\frac{\sum(X_i-\bar{X})^2}{n-1}} < \sqrt{\frac{\sum(X_i-A)^2}{n-1}}$, $A \neq \bar{X}$
2. For normal (symmetric distribution the following holds.

   - Approximately 68.27% of the data values fall within one standard deviation of the mean. i.e. with in $(\bar{X}-S,\ \bar{X}+S)$
   - Approximately 95.45% of the data values fall within two standard deviations of the mean. i.e. with in $(\bar{X}-2S,\ \bar{X}+2S)$
   - Approximately 99.73% of the data values fall within three standard deviations of the mean. i.e. with in $(\bar{X}-3S,\ \bar{X}+3S)$

3. **Chebyshev's Theorem**

   For any data set ,no matter what the pattern of variation, the proportion of the values that fall with in k standard deviations of the mean or $(\bar{X}-kS,\ \bar{X}+kS)$ will be at least $1-\frac{1}{k^2}$, where k is a number greater than 1. i.e. the proportion of items falling beyond k standard deviations of the mean is at most $\frac{1}{k^2}$

   Example: Suppose a distribution has mean 50 and standard deviation 6. What percent of the numbers are:
   a) Between 38 and 62
   b) Between 32 and 68
   c) Less than 38 or more than 62.
   d) Less than 32 or more than 68.

   Solutions:

   a) 38 and 62 are at equal distance from the mean,50 and this distance is 12

$\Rightarrow ks = 12$

$\Rightarrow k = \frac{12}{S} = \frac{12}{6} = 2$

➔ Applying the above theorem at least $(1-\frac{1}{k^2})*100\% = 75\%$ of the numbers lie between 38 and 62.

   b) Similarly done.

c) It is just the complement of a) i.e. at most $\frac{1}{k^2}*100\% = 25\%$ of the numbers lie less than 32 or more than 62.

d) Similarly done.

Example 2:

The average score of a special test of knowledge of wood refinishing has a mean of 53 and standard deviation of 6. Find the range of values in which at least 75% the scores will lie. (Exercise)

4. If the standard deviation of $X_1, X_2, ...X_n \; is \; S$, then the standard deviation of
   a) $X_1 + k,\; X_2 + k, ...X_n + k \;\; will \, also \, be \; S$
   b) $kX_1,\; kX_2, ...kX_n \; would \, be \left|k\right| S$
   c) $a + kX_1,\; a + kX_2, ...a + kX_n \; would \;\; be \left|k\right| \; S$

   Exercise: Verify each of the above relation ship, considering *k* and *a* as constants.

Examples:

1. The mean and standard deviation of n Tetracycline Capsules $X_1, X_2, ...X_n$ are known to be 12 gm and 3 gm respectively. New set of capsules of another drug are obtained by the linear transformation $Y_i = 2X_i - 0.5$ (i = 1, 2, …, n) then what will be the standard deviation of the new set of capsules
2. The mean and the standard deviation of a set of numbers are respectively 500 and 10.
   a. If 10 is added to each of the numbers in the set, then what will be the variance and standard deviation of the new set?
   b. If each of the numbers in the set are multiplied by -5, then what will be the variance and standard deviation of the new set?

Solutions:

1. Using c) above the new standard deviation = $\left|k\right|S = 2*3 = 6$
2. a. They will remain the same.

   b. New standard deviation= $= \left|k\right|S = 5*10 = 50$

Coefficient of Variation (C.V)

- Is defined as the ratio of standard deviation to the mean usually expressed as percents.

$$C.V = \frac{S}{\overline{X}}*100$$

- The distribution having less C.V is said to be less variable or more consistent.

Examples:

1. An analysis of the monthly wages paid (in Birr) to workers in two firms A and B belonging to the same industry gives the following results

Value	Firm A	Firm B
Mean wage	52.5	47.5
Median wage	50.5	45.5
Variance	100	121

In which firm A or B is there greater variability in individual wages?

Solutions:
Calculate coefficient of variation for both firms.

$$C.V_A = \frac{S_A}{\overline{X}_A} * 100 = \frac{10}{52.5} * 100 = 19.05\%$$

$$C.V_B = \frac{S_B}{\overline{X}_B} * 100 = \frac{11}{47.5} * 100 = 23.16\%$$

Since $C.V_A < C.V_B$, in firm B there is greater variability in individual wages.

2. A meteorologist interested in the consistency of temperatures in three cities during a given week collected the following data. The temperatures for the five days of the week in the three cities were

City 1	25	24	23	26	17
City2	22	21	24	22	20
City3	32	27	35	24	28

Which city have the most consistent temperature, based on these data?
(Exercise)

Standard Scores (Z-scores)

- If X is a measurement from a distribution with mean $\overline{X}$ and standard deviation S, then its value in standard units is

$$Z = \frac{X - \mu}{\sigma}, \quad for population$$

$$Z = \frac{X - \overline{X}}{S}, \quad for sample$$

- Z gives the deviations from the mean in units of standard deviation
- Z gives the number of standard deviation a particular observation lie above or below the mean.
- It is used to compare two observations coming from different groups.

Examples:

1. Two sections were given introduction to statistics examinations. The following information was given.

Value	Section 1	Section 2
Mean	78	90
Stan.deviation	6	5

Student A from section 1 scored 90 and student B from section 2 scored 95.Relatively speaking who performed better?

Solutions:

Calculate the standard score of both students.

$$Z_A = \frac{X_A - \overline{X}_1}{S_1} = \frac{90-78}{6} = 2$$

$$Z_B = \frac{X_B - \overline{X}_2}{S_2} = \frac{95-90}{5} = 1$$

➔ Student A performed better relative to his section because the score of student A is two standard deviation above the mean score of his section while, the score of student B is only one standard deviation above the mean score of his section.

2. Two groups of people were trained to perform a certain task and tested to find out which group is faster to learn the task. For the two groups the following information was given:

Value	Group one	Group two
Mean	10.4 min	11.9 min
Stan.dev.	1.2 min	1.3 min

Relatively speaking:

a) Which group is more consistent in its performance
b) Suppose a person A from group one take 9.2 minutes while person B from Group two take 9.3 minutes, who was faster in performing the task? Why?

Solutions:

a) Use coefficient of variation.

$$C.V_1 = \frac{S_1}{\overline{X}_1} * 100 = \frac{1.2}{10.4} * 100 = 11.54\%$$

$$C.V_2 = \frac{S_2}{\overline{X}_2} * 100 = \frac{1.3}{11.9} * 100 = 10.92\%$$

Since $C.V_2 < C.V_1$, group 2 is more consistent.

b) Calculate the standard score of A and B

$$Z_A = \frac{X_A - \overline{X}_1}{S_1} = \frac{9.2 - 10.4}{1.2} = -1$$

$$Z_B = \frac{X_B - \overline{X}_2}{S_2} = \frac{9.3 - 11.9}{1.3} = -2$$

➔Child B is faster because the time taken by child B is two standard deviation shorter than the average time taken by group 2 while, the time taken by child A is only one standard deviation shorter than the average time taken by group 1.

4.3 Moments

- If X is a variable that assume the values $X_1, X_2, \ldots, X_n$ then

1. The r^{th} moment is defined as:

$$\overline{X}^r = \frac{X_1^{\,r} + X_2^{\,r} + \ldots + X_n^{\,r}}{n}$$

$$= \frac{\sum_{i=1}^{n} X_i^{\,r}}{n}$$

- For the case of frequency distribution this is expressed as:

$$\overline{X}^r = \frac{\sum_{i=1}^{k} f_i X_i^{\,r}}{n}$$

- If $r = 1$, it is the simple arithmetic mean, this is called the first moment.

2. The r^{th} moment about the mean (the r^{th} central moment)

- Denoted by M_r and defined as:

$$M_r = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^r}{n} = \frac{(n-1)}{n} \frac{\sum_{i=1}^{n} (X_i - \overline{X})^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r = \frac{\sum_{i=1}^{k} f_i (X_i - \overline{X})^r}{n}$$

- If $r = 2$, it is population variance, this is called the second central moment. If we assume $n - 1 \approx n$,it is also the sample variance.

3. The r^{th} moment about any number A is defined as:

- Denoted by $M_r^{'}$ and

$$M_r' = \frac{\sum_{i=1}^{n}(X_i - A)^r}{n} = \frac{(n-1)}{n}\frac{\sum_{i=1}^{n}(X_i - A)^r}{n-1}$$

- For the case of frequency distribution this is expressed as:

$$M_r' = \frac{\sum_{i=1}^{k} f_i(X_i - A)^r}{n}$$

Example:

1. Find the first two moments for the following set of numbers 2, 3, 7
2. Find the first three central moments of the numbers in problem 1
3. Find the third moment about the number 3 of the numbers in problem 1.

Solutions:

1. Use the r[th] moment formula.

$$\overline{X}^r = \frac{\sum_{i=1}^{n} X_i^r}{n}$$

$$\Rightarrow \overline{X}^1 = \frac{2+3+7}{3} = 4 = \overline{X}$$

$$\overline{X}^2 = \frac{2^2+3^2+7^2}{3} = 20.67$$

2. Use the r[th] central moment formula.

$$M_r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^r}{n}$$

$$\Rightarrow M_1 = \frac{(2-4)+(3-4)+(7-4)}{3} = 0$$

$$M_2 = \frac{(2-4)^2+(3-4)^2+(7-4)^2}{3} = 4.67$$

$$M_3 = \frac{(2-4)^3+(3-4)^3+(7-4)^3}{3} = 6$$

3. Use the r[th] moment about A.

$$M_r = \frac{\sum_{i=1}^{n}(X_i - A)^r}{n}$$

$$\Rightarrow M_3' = \frac{(2-3)^3 + (3-3)^3 + (7-3)^3}{3} = 21$$

4.4 Skewness

- Skewness is the degree of asymmetry or departure from symmetry of a distribution.
- A skewed frequency distribution is one that is not symmetrical.
- Skewness is concerned with the shape of the curve not size.
- If the frequency curve (smoothed frequency polygon) of a distribution has a longer tail to the right of the central maximum than to the left, the distribution is said to be skewed to the right or said to have positive skewness. If it has a longer tail to the left of the central maximum than to the right, it is said to be skewed to the left or said to have negative skewness.
- For moderately skewed distribution, the following relation holds among the three commonly used measures of central tendency.

$$Mean - Mode = 3*(Mean - Medi)$$

Measures of Skewness

- Denoted by α_3
- There are various measures of skewness.

1. The Pearsonian coefficient of skewness

$$\alpha_3 = \frac{Mean - Mode}{Standard\ deviation\ S} = \frac{\bar{X} - \hat{X}}{S}$$

2. The Bowley's coefficient of skewness (coefficient of skewness based on quartiles)

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1}$$

3. The moment coefficient of skewness

$$\alpha_3 = \frac{M_3}{M_2^{3/2}} = \frac{M_3}{(\sigma^2)^{3/2}} = \frac{M_3}{\sigma^3}, \text{ Where } \sigma \text{ is the population standard deviati}$$

The shape of the curve is determined by the value of α_3

- *If* $\alpha_3 > 0$ *then the distribution is positively skewed.*
- *If* $\alpha_3 = 0$ *then the distribution is symmetric.*
- *If* $\alpha_3 < 0$ *then the distribution is negatively skewed.*

Remark:

- In a positively skewed distribution, smaller observations are more frequent than larger observations. i.e. the majority of the observations have a value below an average.
- In a negatively skewed distribution, smaller observations are less frequent than larger observations. i.e. the majority of the observations have a value above an average.

Examples:

1. Suppose the mean, the mode, and the standard deviation of a certain distribution are 32, 30.5 and 10 respectively. What is the shape of the curve representing the distribution?

   Solutions:

   Use the Pearsonian coefficient of skewness

$$\alpha_3 = \frac{Mean - Mode}{S\tan dard\ deviation} = \frac{32 - 30.5}{10} = 0.15$$

$$\alpha_3 > 0 \Rightarrow The\ distribution\ is\ positively\ skewed.$$

2. In a frequency distribution, the coefficient of skewness based on the quartiles is given to be 0.5. If the sum of the upper and lower quartile is 28 and the median is 11, find the values of the upper and lower quartiles.

   Solutions:

Given: $\alpha_3 = 0.5, \quad \tilde{X} = Q_2 = 11$ Required: Q_1, Q_3

$$Q_1 + Q_3 = 28 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (*)$$

$$\alpha_3 = \frac{(Q_3 - Q_2) - (Q_2 - Q_1)}{Q_3 - Q_1} = \frac{Q_3 + Q_1 - 2Q_2}{Q_3 - Q_1} = 0.5$$

Substituting the given values, one can obtain the following

$$Q_3 - Q_1 = 12 \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (**)$$

Solving (*) and (**) at the same time we obtain the following values

$Q_1 = 8$ *and* $Q_3 = 20$

3. Some characteristics of annually family income distribution (in Birr) in two regions is as follows:

Region	Mean	Median	Standard Deviation
A	6250	5100	960
B	6980	5500	940

   a) Calculate coefficient of skewness for each region
   b) For which region is, the income distribution more skewed. Give your interpretation for this Region
   c) For which region is the income more consistent?

Solutions: (**exercise)**

4. For a moderately skewed frequency distribution, the mean is 10 and the median is 8.5. If the coefficient of variation is 20%, find the Pearsonian coefficient of skewness and the probable mode of the distribution. (**exercise)**

5. The sum of fifteen observations, whose mode is 8, was found to be 150 with coefficient of variation of 20%
   (a) Calculate the Pearsonian coefficient of skewness and give appropriate conclusion.
   (b) Are smaller values more or less frequent than bigger values for this distribution?
   (c) If a constant *k* was added on each observation, what will be the new Pearsonian coefficient of skewness? Show your steps. What do you conclude from this?
   (**Exercise**)

4.5 Kurtosis

Kurtosis is the degree of peakdness of a distribution, usually taken relative to a normal distribution. A distribution having relatively high peak is called *leptokurtic*. If a curve representing a distribution is flat topped, it is called *platykurtic*. The normal distribution which is not very high peaked or flat topped is called *mesokurtic*.

Measures of kurtosis
The moment coefficient of kurtosis:

- Denoted by α_4 and given by

$$\alpha_4 = \frac{M_4}{M_2{}^2} = \frac{M_4}{\sigma^4}$$

$$Where: M_4 \text{ is the fourth moment about the mean.}$$
$$M_2 \text{ is the second moment about the mean.}$$
$$\sigma \text{ is the population standard deviation.}$$

The peakdness depends on the value of α_4.

$$If\ \alpha_4 > 3 then\ the\ curve\ is\ leptokurtic.$$
$$If\ \alpha_4 = 3 then\ the\ curve\ is\ mesokurtic.$$
$$If\ \alpha_4 < 3 then\ the\ curve\ is\ platykurtic.$$

Examples:

1. If the first four central moments of a distribution are:

$$M_1 = 0,\ M_2 = 16,\ M_3 = -60,\ M_4 = 162$$

   a) Compute a measure of skewness
   b) Compute a measure of kurtosis and give your interpretation.

Solutions:

a) $$\alpha_3 = \frac{M_3}{M_2^{3/2}} = \frac{-60}{16^{3/2}} = -0.94 < 0$$

$$\Rightarrow \textit{The distribution is negatively skewed.}$$

b) $$\alpha_4 = \frac{M_4}{M_2^{2}} = \frac{162}{16^{2}} = 0.6 < 3$$

$$\Rightarrow \textit{The curve is platykurtic.}$$

2. The median and the mode of a mesokurtic distribution are 32 and 34 respectively. The 4^{th} moment about the mean is 243. Compute the Pearsonian coefficient of skewness and identify the type of skewness. Assume (n-1 = n) (**exercise)**.

3. If the standard deviation of a symmetric distribution is 10, what should be the value of the fourth moment so that the distribution is mesokurtic?
   Solutions (**exercise)**.

CHAPTER 5
5. ELEMENTARY PROBABILITY

5.1 Introduction

- Probability theory is the foundation upon which the logic of inference is built.
- It helps us to cope up with uncertainty.
- In general, probability is the chance of an outcome of an experiment. It is the measure of how likely an outcome is to occur.

5.2 Definitions of some probability terms

1. **Experiment**: Any process of observation or measurement or any process which generates well defined outcome.
2. **Probability Experiment**: It is an experiment that can be repeated any number of times under similar conditions and it is possible to enumerate the total number of outcomes without predicting an individual out come. It is also called random experiment.

Example: If a fair die is rolled once it is possible to list all the possible outcomes i.e.1, 2, 3, 4, 5, 6 but it is not possible to predict which outcome will occur.

3. **Outcome**: The result of a single trial of a random experiment
4. **Sample Space**: Set of all possible outcomes of a probability experiment
5. **Event**: It is a subset of sample space. It is a statement about one or more outcomes of a random experiment. They are denoted by capital letters.

Example: Considering the above experiment let A be the event of odd numbers, B be the event of even numbers, and C be the event of number 8.

$\Rightarrow A = \{1,3,5\}$

$B = \{2,4,6\}$

$C = \{\ \}\ or\ empty\ space\ or\ \ impossible\ event$

Remark:

If S (sample space) has n members then there are exactly 2^n subsets or events.

6. **Equally Likely Events**: Events which have the same chance of occurring.
7. **Complement of an Event:** the complement of an event A means non-occurrence of A and is denoted by $A', or\ A^c, or\ \overline{A}$ contains those points of the sample space which don't belong to A.
8. **Elementary Event**: an event having only a single element or sample point.
9. **Mutually Exclusive Events**: Two events which cannot happen at the same time.

10. **Independent Events**: Two events are independent if the occurrence of one does not affect the probability of the other occurring.

11. **Dependent Events**: Two events are dependent if the first event affects the outcome or occurrence of the second event in a way the probability is changed.

Example: .What is the sample space for the following experiment

a) Toss a die one time.
b) Toss a coin two times.
c) A light bulb is manufactured. It is tested for its life length by time.

Solution

a) S={1,2,3,4,5,6}
b) S={(HH),(HT),(TH),(TT)}
c) S={t /t≥0}
   - Sample space can be
     - Countable (finite or infinite)
     - Uncountable

5.3 Counting Rules

In order to calculate probabilities, we have to know

- The number of elements of an event
- The number of elements of the sample space.

That is in order to judge what is **probable**, we have to know what is **possible.**

- In order to determine the number of outcomes, one can use several rules of counting.
  - The addition rule
  - The multiplication rule
  - Permutation rule
  - Combination rule

- To list the outcomes of the sequence of events, a useful device called **tree diagram is** used.

The addition rule

Suppose that the 1st procedure designed by 1 can be performed in n1 ways. Assume that 2nd procedure designed by 2 can be performed in n2 ways.
suppose further more that, it is not possible that both procedures 1 and 2 are performed together then the number of ways in which we can perform 1or 2 procedure is n_1+n_2 ways, and also if we have another procedure that is designed by k with possible way of n_k we can conclude that there is $n_1+n_2+\ldots+n_k$ possible ways.

Example: suppose we planning a trip and are deciding by bus and train transportation. If there are 3 bus routes and 2 train routes to go from A to B. find the available routes for the trip.

Solution:

There are 3+2 =5 routes for someone to go from A to B.

The Multiplication Rule:

If a choice consists of k steps of which the first can be made in n_1 ways, the second can be made in n_2 ways… the k^{th} can be made in n_k ways, then the whole choice can be made in $(n_1 * n_2 * ..... * n_k)\ ways.$

Example 1

An air line has 6 flights from A to B, and 7 flights from B to C per day. If the flights are to be made on separate days, in how many different ways can the airline offer from A to C?

Solution: In operation 1 there are 6 flights from A to B, 7 flights are available to make flight from B to C. Altogether there are 6*7 = 42 possible flights from A to C.

Example2

suppose that in a medical study patients are classified according to their blood type as A, B , AB, and O; according to their RH factors as + or - and according to their blood pressure as high, normal or low ,then in how many different ways can a patient be classified ?

Solution

The 1st classification done in 4 ways, the 2nd in 2 ways, and the 3rd in 3 ways. Thus patient can be classified in 4*2*3 = 24 different ways.

Example 3

The digits 0, 1, 2, 3, and 4 are to be used in 4 digit identification card. How many different cards are possible if

a) Repetitions are permitted.
b) Repetitions are not permitted.

Solutions

a)

1st digit	2nd digit	3rd digit	4th digit
5	5	5	5

There are four steps

1. Selecting the 1st digit, this can be made in 5 ways.
2. Selecting the 2nd digit, this can be made in 5 ways.
3. Selecting the 3rd digit, this can be made in 5 ways.
4. Selecting the 4th digit, this can be made in 5 ways.

$\Rightarrow 5*5*5*5 = 625$ *different cards are possible.*

b)

1st digit	2nd digit	3rd digit	4th digit
5	4	3	2

There are four steps

5. Selecting the 1st digit, this can be made in 5 ways.
6. Selecting the 2nd digit, this can be made in 4 ways.
7. Selecting the 3rd digit, this can be made in 3 ways.
8. Selecting the 4th digit, this can be made in 2 ways.

$\Rightarrow 5*4*3*2=120$ *different cards are possible.*

Permutation

An arrangement of n objects in a specified order is called permutation of the objects.

Permutation Rules:

1. The number of permutations of *n* distinct objects taken all together is n!

   Where $n!=n*(n-1)*(n-2)*.....*3*2*1$

   ${}_nP_n=\frac{n!}{(n-n)!}=\frac{n!}{0!}$=n!. In definition 0! = 1! = 1

2. The arrangement of n objects in a specified order using r objects at a time is called the permutation of n objects taken r objects at a time. It is written as ${}_nP_r$ and the formula is

$$ {}_nP_r=\frac{n!}{(n-r)!}$$

3. The number of permutations of n objects in which k_1 are alike k_2 are alike ---- etc is

$$ {}_nP_r=\frac{n!}{k_1!*k_2*...*k_n}$$

Example:

1. Suppose we have a letters A,B, C, D
   a) How many permutations are there taking all the four?
   b) How many permutations are there two letters at a time?
2. How many different permutations can be made from the letters in the word "CORRECTION"?

Solutions:

1.

a)

Here $n=4,$ *there are four disnict object*

$\Rightarrow$ *There are* $4!=24$ *permutations.*

b)

Here $n=4,\quad r=2$

$\Rightarrow$ *There are* ${}_4P_2=\frac{4!}{(4-2)!}=\frac{24}{2}=12$ *permutations.*

2.

$$Here\ n = 10$$

$$Of\ \ which\ \ 2\ are\ C,\ \ 2\ are\ O,\ \ 2\ are\ R\ ,1E,1T,1I,1N$$

$$\Rightarrow K_1 = 2,\ k_2 = 2,\ k_3 = 2,\ \ k_4 = k_5 = k_6 = k_7 = 1$$

$$U\sin g\ the\ 3^{rd}\ \ rule\ of\ \ permutation, there\ are$$

$$\frac{10!}{2!*2!*2!*1!*1!*1!*1!} = 453600\ permutations.$$

Exercises:

1. Six different statistics books, seven different physics books, and 3 different Economics books are arranged on a shelf. How many different arrangements are possible if;
   i. The books in each particular subject must all stand together
   ii. Only the statistics books must stand together
2. If the permutation of the word WHITE is selected at random, how many of the permutations
   i. Begins with a consonant?
   ii. Ends with a vowel?
   iii. Has a consonant and vowels alternating?

Combination

A selection of objects without regard to order is called combination.

Example: Given the letters A, B, C, and D list the permutation and combination for selecting two letters.

Solutions:

Permutation

AB	BA	CA	DA
AC	BC	CB	DB
AD	BD	CD	DC

Combination

AB	BC
AC	BD
AD	DC

Note that in permutation AB is different from BA. But in combination AB is the same as BA.

Combination Rule

The number of combinations of r objects selected from n objects is denoted by ${}_nC_r\ or\ \binom{n}{r}$ and is given by the formula:

$$\binom{n}{r} = \frac{n!}{(n-r)!*r!}$$

Examples:

1. In how many ways a committee of 5 people be chosen out of 9 people?

Solutions:

$$n = 9\ ,\ \ r = 5$$

$$\binom{n}{r} = \frac{n!}{(n-r)!*r!} = \frac{9!}{4!*5!} = 126\,ways$$

2. Among 15 clocks there are two defectives .In how many ways can an inspector chose three of the clocks for inspection so that:
   a) There is no restriction.
   b) None of the defective clock is included.
   c) Only one of the defective clocks is included.
   d) Two of the defective clock is included.

Solutions:

$$n = 15 \ \ of\ which\ \ 2\ are\ defective\ and\ \ 13\ are\ non-defective.$$
$$r = 3$$

a) If there is no restriction select three clocks from 15 clocks and this can be done in :

$$n = 15\ ,\ \ r = 3$$

$$\binom{n}{r} = \frac{n!}{(n-r)!*r!} = \frac{15!}{12!*3!} = 455\,way$$

b) None of the defective clocks is included.
This is equivalent to zero defective and three non defective, which can be done in:

$$\binom{2}{0} * \binom{13}{3} = 286\,ways.$$

c) Only one of the defective clocks is included.
This is equivalent to one defective and two non defective, which can be done in:

$$\binom{2}{1} * \binom{13}{2} = 156 ways.$$

d) Two of the defective clock is included.
This is equivalent to two defective and one non defective, which can be done in:

$$\binom{2}{2} * \binom{13}{1} = 13 \ ways.$$

Exercises:

1. Out of 5 Mathematician and 7 Statistician a committee consisting of 2 Mathematician and 3 Statistician is to be formed. In how many ways this can be done if
   a) There is no restriction
   b) One particular Statistician should be included
   c) Two particular Mathematicians can not be included on the committee.
2. If 3 books are picked at random from a shelf containing 5 novels, 3 books of poems, and a dictionary, in how many ways this can be done if
   a) There is no restriction.
   b) The dictionary is selected?
   c) 2 novels and 1 book of poems are selected?

5.4 Approaches to measuring Probability

There are four different conceptual approaches to the study of probability theory. These are:

- The classical approach.
- The relative frequency approach.
- The axiomatic approach.
- The subjective approach.

The classical approach

This approach is used when:

- All outcomes are equally likely.
- Total number of outcome is finite, say N.

Definition: If a random experiment with N equally likely outcomes is conducted and out of these N_A outcomes are favorable to the event A, then the probability that event A occur denoted $P(A)$ is defined as:

$$P(A) = \frac{N_A}{N} = \frac{No.of\ outcomes\ favourable to\ A}{Total\ number\ of\ \text{outcomes}} = \frac{n(A)}{n(S)}$$

Examples:

1. A fair die is tossed once. What is the probability of getting
   a) Number 4?
   b) An odd number?
   c) An even number?
   d) Number 8?

Solutions:

First identify the sample space, say S

$$S = \{1,2,3,4,5,6\}$$

$$\Rightarrow N = n(S) = 6$$

a) Let A be the event of number 4

$$A = \{4\}$$

$$\Rightarrow N_A = n(A) = 1$$

$$P(A) = \frac{n(A)}{n(S)} = 1/6$$

b) Let A be the event of odd numbers

$$A = \{1,3,5\}$$

$$\Rightarrow N_A = n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = 3/6 = 0.5$$

c) Let A be the event of even numbers

$$A = \{2,4,6\}$$

$$\Rightarrow N_A = n(A) = 3$$

$$P(A) = \frac{n(A)}{n(S)} = 3/6 = 0.5$$

d) Let A be the event of number 8

$$A = \text{Ø}$$

$$\Rightarrow N_A = n(A) = 0$$

$$P(A) = \frac{n(A)}{n(S)} = 0/6 = 0$$

2. A box of 80 candles consists of 30 defective and 50 non defective candles. If 10 of this candles are selected at random, what is the probability
   a) All will be defective.
   b) 6 will be non defective

c) All will be non defective

Solutions:

$$Total\ selection = \binom{80}{10} = N = n(S)$$

a) Let A be the event that all will be defective.

$$Total\ way\ in\ which\ A\ occur = \binom{30}{10} * \binom{50}{0} = N_A = n(A)$$

$$\Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{\binom{30}{10} * \binom{50}{0}}{\binom{80}{10}} = 0.00001825$$

b) Let A be the event that 6 will be non defective.

$$Total\ way\ in\ which\ A\ occur = \binom{30}{4} * \binom{50}{6} = N_A = n(A)$$

$$\Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{\binom{30}{4} * \binom{50}{6}}{\binom{80}{10}} = 0.265$$

c) Let A be the event that all will be non defective.

$$Total\ way\ in\ which\ A\ occur = \binom{30}{0} * \binom{50}{10} = N_A = n(A)$$

$$\Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{\binom{30}{0} * \binom{50}{10}}{\binom{80}{10}} = 0.0\,0\,6\,2\,4$$

Exercises:

1. What is the probability that a waitress will refuse to serve alcoholic beverages to only three minors if she randomly checks the I.D's of five students from among ten students of which four are not of legal age?

2. If 3 books are picked at random from a shelf containing 5 novels, 3 books of poems, and a dictionary, what is the probability that
   a) The dictionary is selected?
   b) 2 novels and 1 book of poems are selected?

Short coming of the classical approach:

This approach is not applicable when:

- The total number of outcomes is infinite.
- Outcomes are not equally likely.

The Frequentist Approach

This is based on the relative frequencies of outcomes belonging to an event.

Definition: The probability of an event A is the proportion of outcomes favorable to A in the long run when the experiment is repeated under same condition.

$$P(A) = \lim_{N\to\infty} \frac{N_A}{N}$$

Example: If records show that 60 out of 100,000 bulbs produced are defective. What is the probability of a newly produced bulb to be defective?

Solution:

Let A be the event that the newly produced bulb is defective.

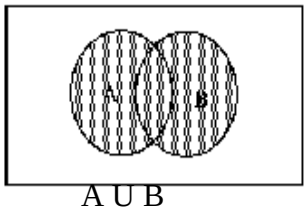
$$P(A) = \lim_{N\to\infty} \frac{N_A}{N} = \frac{60}{100{,}000} = 0.0006$$

Axiomatic Approach:

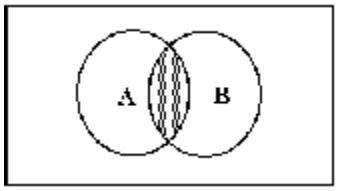
Let E be a random experiment and S be a sample space associated with E. With each event A a real number called the probability of A satisfies the following properties called axioms of probability or postulates of probability.

1. $P(A) \geq 0$
2. $P(S) = 1,\ S \ is the sure \ \ event.$
3. If A and B are mutually exclusive events, the probability that one or the other occur equals the sum of the two probabilities. i. e.
   $$P(A \cup B) = P(A) + P(B)$$
4. $P(A') = 1 - P(A)$
5. $0 \leq P(A) \leq 1$
6. P (ø) =0, ø is the impossible event.

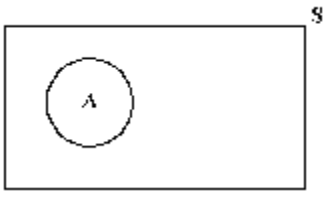
Remark: Venn-diagrams can be used to solve probability problems.



A U B



A n B



A

In general $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Conditional probability and Independency

Conditional Events: If the occurrence of one event has an effect on the next occurrence of the other event then the two events are conditional or dependant events.

Example: Suppose we have two red and three white balls in a bag

1. Draw a ball with replacement

   Let A= the event that the first draw is red ➔ $p(A) = \frac{2}{5}$

   B= the event that the second draw is red ➔ $p(B) = \frac{2}{5}$

   A and B are independent.

2. Draw a ball without replacement

   Let A= the event that the first draw is red ➔ $p(A) = \frac{2}{5}$

   B= the event that the second draw is red ➔ $p(B) = ?$

   This is conditional.

Let B= the event that the second draw is red given that the first draw is red ➔ $p(B) = 1/4$

Conditional probability of an event

The conditional probability of an event A given that B has already occurred, denoted $p(A/B)$ is

$$p(A/B) = \frac{p(A \cap B)}{p(B)}, \qquad p(B) \neq 0$$

Remark: (1) $p(A'/B) = 1 - p(A/B)$

(2) $p(B'/A) = 1 - p(B/A)$

Examples

1. For a student enrolling at freshman at certain university the probability is 0.25 that he/she will get scholarship and 0.75 that he/she will graduate. If the probability is 0.2 that he/she will get scholarship and will also graduate. What is the probability that a student who get a scholarship graduate?

   Solution: Let A= the event that a student will get a scholarship

   B= the event that a student will graduate

   $given \quad p(A) = 0.25, \qquad p(B) = 0.75, \qquad p(A \cap B) = 0.20$

   $\text{Required} \quad p(B/A)$

   $$p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{0.20}{0.25} = 0.80$$

2. If the probability that a research project will be well planned is 0.60 and the probability that it will be well planned and well executed is 0.54, what is the probability that it will be well executed given that it is well planned?

Solution; Let A= the event that a research project will be well Planned

B= the event that a research project will be well Executed

$$given \quad p(A) = 0.60, \qquad p(A \cap B) = 0.54$$

$$\text{Re}\, quired \quad p(B/A)$$

$$p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{0.54}{0.60} = 0.90$$

3. A lot consists of 20 defectives and 80 non-defective items from which two items are chosen without replacement. Events A & B are defined as A = {the first item chosen is defective}, B = {the second item chosen is defective}
   a. What is the probability that both items are defective?
   b. What is the probability that the second item is defective?

Solution; Exercise

Note; for any two events A and B the following relation holds.

$$p(B) = p(B/A).p(A) + p(B/A').p(A')$$

Probability of Independent Events

Two events A and B are independent if and only if $p(A \cap B) = p(A).p(B)$

Here $p(A/B) = p(A), \qquad P(B/A) = p(B)$

Example; A box contains four black and six white balls. What is the probability of getting two black balls in drawing one after the other under the following conditions?

a. The first ball drawn is not replaced
b. The first ball drawn is replaced

Solution; Let A= first drawn ball is black

B= second drawn is black

Required $p(A \cap B)$

a. $p(A \cap B) = p(B/A).p(A) = (4/10)(3/9) = 2/15$
b. $p(A \cap B) = p(A).p(B) = (4/10)(4/10) = 4/25$

CHAPTER 6

6. RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

6.1 Definitions of random variable and probability distributions

Definition: *A random variable* is a numerical description of the outcomes of the experiment or a numerical valued function defined on sample space, usually denoted by capital letters.

Example: If X is a random variable, then it is a function from the elements of the sample space to the set of real numbers. i.e.

X is a function X: S $\rightarrow$ R

➔A random variable takes a possible outcome and assigns a number to it.

Example: Flip a coin three times, let X be the number of heads in three tosses.

$$\Rightarrow S = \{(HHH),(HHT),(HTH),(HTT),(THH),(THT),(TTH),(TTT)\}$$

$$\Rightarrow X(HHH) = 3, \quad X(HHT) = X(HTH) = X(THH) = 2,$$
$$X(HTT) = X(THT) = X(TTH) = 1$$
$$X(TTT) = 0$$

X = {0, 1, 2, 3}

➔X assumes a specific number of values with some probabilities.

Random variables are of two types:

1. *Discrete random variable*: are variables which can assume only a specific number of values. They have values that can be counted

Examples:

- Toss coin n times and count the number of heads.
- Number of children in a family.
- Number of car accidents per week.
- Number of defective items in a given company.
- Number of bacteria per two cubic centimeter of water.

2. *Continuous random variable*: are variables that can assume all values between any two given values.

Examples:

- Height of students at certain college.
- Mark of a student.
- Life time of light bulbs.
- Length of time required to complete a given training.

Definition: *a probability distribution* consists of a value a random variable can assume and the corresponding probabilities of the values.

Example: Consider the experiment of tossing a coin three times. Let X is the number of heads. Construct the probability distribution of X.

Solution:

- First identify the possible value that X can assume.
- Calculate the probability of each possible distinct value of X and express X in the form of frequency distribution.

$X = x$	0	1	2	3
$P(X = x)$	$1/8$	$3/8$	$3/8$	$1/8$

Probability distribution is denoted by P for discrete and by f for continuous random variable.

Properties of Probability Distribution:

1. $P(x) \geq 0, \quad if \quad X \quad is \ discrete.$
   $f(x) \geq 0, \ if \ X \ is \ continuous.$

2. $\sum_x P(X = x) = 1, \quad if \quad X \quad is \quad discrete.$
   $\int_x f(x)dx = 1, \quad if \ is c \ o \ n \ t \ i \ n \ u \ o$

Note:

1. If X is a continuous random variable then

$$P(a < X < b) = \int_a^b f(x)dx$$

2. Probability of a fixed value of a continuous random variable is zero.

$$\Rightarrow P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

3. If X is discrete random variable the

$$P(a < X < b) = \sum_{x=a+1}^{b-1} P(x)$$
$$P(a \leq X < b) = \sum_{x=a}^{b-1} p(x)$$
$$P(a < X \leq b) = \sum_{x=a+1}^{b} P(x)$$
$$P(a \leq X \leq b) = \sum_{x=a}^{b} P(x)$$

4. Probability means *area* for continuous random variable.

6.2 Introduction to expectation

Definition:

1. Let a discrete random variable X assume the values X_1, X_2, …, X_n with the probabilities $P(X_1)$, $P(X_2)$, ….,$P(X_n)$ respectively. Then the expected value of X ,denoted as E(X) is defined as:

$$E(X) = X_1P(X_1) + X_2P(X_2) + ... + X_nP(X_n)$$
$$= \sum_{i=1}^{n} X_iP(X_i)$$

2. Let X be a continuous random variable assuming the values in the interval (a, b) such that $\int_a^b f(x)dx = 1$,then

$$E(X) = \int_a^b x\, f(x)dx$$

Examples:

1. What is the expected value of a random variable X obtained by tossing a coin three times where is the number of heads

Solution:

First construct the probability distribution of X

$X = x$	0	1	2	3
• $P(X = x)$	• $1/8$	• $3/8$	• $3/8$	• $1/8$

$$\Rightarrow E(X) = X_1P(X_1) + X_2P(X_2) + ... + X_nP(X_n)$$
$$= 0*1/8 + 1*3/8 + ... + 2*1/8$$
$$= 1.5$$

2. Suppose a charity organization is mailing printed return-address stickers to over one million homes in the Ethiopia. Each recipient is asked to donate$1, $2, $5, $10, $15, or $20. Based on past experience, the amount a person donates is believed to follow the following probability distribution:

$X = x$	\$1	\$2	\$5	\$10	\$15	\$20
$P(X = x)$	0.1	0.2	0.3	0.2	0.15	0.05

What is expected that an average donor to contribute?

Solution:

$X = x$	\$1	\$2	\$5	\$10	\$15	\$20	Total
$P(X = x)$	0.1	0.2	0.3	0.2	0.15	0.05	1
$xP(X = x)$	0.1	0.4	1.5	2	2.25	1	7.25

$$\Rightarrow E(X) = \sum_{i=1}^{6} x_i P(X = x_i) = \$7.25$$

Mean and Variance of a random variable

Let X is given random variable.

1. The expected value of X is its mean $\Rightarrow Mean of\ X = E(X)$
2. The variance of X is given by:

$$Variance of\ X = \text{var}(X) = E(X^2) - [E(X)]^2$$

Where:

$$E(X^2) = \sum_{i=1}^{n} x_i^{\,2} P(X = x_i)\ ,\quad if\ \ X\ \ is\ discrete$$
$$= \int_x x^2 f(x)dx\ ,\qquad if\ \ X\ is\ continuous.$$

Examples:

1. Find the mean and the variance of a random variable X in example 2 above.

Solutions:

$X = x$	\$1	\$2	\$5	\$10	\$15	\$20	Total

$P(X=x)$	0.1	0.2	0.3	0.2	0.15	0.05	1
$xP(X=x)$	0.1	0.4	1.5	2	2.25	1	7.25
$x^2P(X=x)$	0.1	0.8	7.5	20	33.75	20	82.15

$\Rightarrow E(X)=7.25$

$Var(X)=E(X^2)-[E(X)]^2=82.15-7.25^2=29.59$

2. Two dice are rolled. Let X is a random variable denoting the sum of the numbers on the two dice.
   i) Give the probability distribution of X
   ii) Compute the expected value of X and its variance

There are some general rules for mathematical expectation.

Let X and Y are random variables and *k* is a constant.

RULE 1 $E(k)=k$

RULE 2 $Var(k)=0$

RULE 3 $E(kX)=kE(X)$

RULE 4 $Var(kX)=k^2Var(X)$

RULE 5 $E(X+Y)=E(X)+E(Y)$

6.3 Common Discrete Probability Distributions

1. Binomial Distribution

A binomial experiment is a probability experiment that satisfies the following four requirements called assumptions of a binomial distribution.

1. The experiment consists of n identical trials.
2. Each trial has only one of the two possible mutually exclusive outcomes, success or a failure.
3. The probability of each outcome does not change from trial to trial, and

4. The trials are independent, thus we must sample with replacement.

Examples of binomial experiments

- Tossing a coin 20 times to see how many tails occur.
- Asking 200 people if they watch BBC news.
- Registering a newly produced product as defective or non defective.
- Asking 100 people if they favor the ruling party.
- Rolling a die to see if a 5 appears.

Definition: The outcomes of the binomial experiment and the corresponding probabilities of these outcomes are called **Binomial Distribution**.

$$Let\ P = the\ probability\ of\ success$$

$$q = 1 - p = the\ probability\ of\ \ failure\ on\ any\ given\ trial$$

Then the probability of getting x successes in n trials becomes:

$$P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,2,\ldots,n$$

And this is sometimes written as:

$$X \sim Bin(n, p)$$

When using the binomial formula to solve problems, we have to identify three things:

- The number of trials (n)
- The probability of a success on any one trial (p) and
- The number of successes desired (X).

Examples:

1. What is the probability of getting three heads by tossing a fair coin four times?

Solution:

Let X be the number of heads in tossing a fair coin four times

$$X \sim Bin(n = 4, p = 0.50)$$

$$\Rightarrow P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,2,3,4.$$

$$\Rightarrow P(X = 3) = \binom{4}{3} 0.5^4 = 0.25$$

2. **S**uppose that an examination consists of six true and false questions, and assume that a student has no knowledge of the subject matter. The probability that the student will guess the correct answer to the first question is 30%. Likewise, the probability of guessing each of the remaining questions correctly is also 30%.
   a) What is the probability of getting more than three correct answers?
   b) What is the probability of getting at least two correct answers?
   c) What is the probability of getting at most three correct answers?
   d) What is the probability of getting less than five correct answers?

Solution

Let X = the number of correct answers that the student gets.

$$X \sim Bin(n = 6, p = 0.30)$$

a) $P(X > 3) = ?$

$$\Rightarrow P(X = x) = \binom{n}{x} p^x q^{n-x}, \quad x = 0,1,2,..6$$

$$= \binom{6}{x} 0.3^x 0.7^{6-x}$$

$$\Rightarrow P(X > 3) = P(X = 4) + P(X = 5) + P(X = 6)$$

$$= 0.060 + 0.010 + 0.001$$

$$= 0.071$$

Thus, we may conclude that if 30% of the exam questions are answered by guessing, the probability is 0.071 (or 7.1%) that more than four of the questions are answered correctly by the student.

b) $P(X \geq 2) = ?$

$$P(X \geq 2) = P(X = 2) + P(X = 3) + P(X = 4) + P(X = 5) + P(X = 6)$$

$$= 0.324 + 0.185 + 0.060 + 0.010 + 0.001$$

$$= 0.58$$

c) $P(X \leq 3) = ?$

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)$$

$$= 0.118 + 0.303 + 0.324 + 0.185$$

$$= 0.93$$

d) $P(X < 5) = ?$

$$P(X < 5) = 1 - P(X \geq 5)$$

$$= 1 - \{P(X = 5) + P(X = 6)\}$$

$$= 1 - (0.010 + 0.001)$$

$$= 0.989$$

Exercises:

1. Suppose that 4% of all TVs made by A&B Company in 2000 are defective. If eight of these TVs are randomly selected from across the country and tested, what is the probability that *exactly* three of them are defective? Assume that each TV is made independently of the others.
2. An allergist claims that 45% of the patients she tests are allergic to some type of weed. What is the probability that
   a) Exactly 3 of her next 4 patients are allergic to weeds?
   b) None of her next 4 patients are allergic to weeds?
3. Explain why the following experiments are not Binomial

   - Rolling a die until a 6 appears.
   - Asking 20 people how old they are.
   - Drawing 5 cards from a deck for a poker hand.

Remark: If X is a binomial random variable with parameters n and p then

$$E(X) = np \quad , \qquad Var(X) = npq$$

2. Poisson Distribution

- A random variable X is said to have a Poisson distribution if its probability distribution is given by:

$$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0,1,2,.....$$

$$Where \ \ \lambda = the \ average \, number.$$

- The Poisson distribution depends *only* on the average number of occurrences per unit time of space.
- The Poisson distribution is used as a distribution of rare events, such as:
  - Number of misprints.
  - Natural disasters like earth quake.
  - Accidents.
  - Hereditary.
  - Arrivals

- The process that gives rise to such events are called Poisson process.

Examples:

1. If 1.6 accidents can be expected an intersection on any given day, what is the probability that there will be 3 accidents on any given day?

Solution; Let X =the number of accidents, $\lambda = 1.6$

$$X = poisson(1.6) \Rightarrow p(X = x) = \frac{1.6^x e^{-1.6}}{x!}$$

$$p(X = 3) = \frac{1.6^3 e^{-1.6}}{3!} = 0.1380$$

Exercise

2. On the average, five smokers pass a certain street corners every ten minutes, what is the probability that during a given 10minutes the number of smokers passing will be
   a. 6 or fewer
   b. 7 or more
   c. Exactly 8…….

If X is a Poisson random variable with parameters λ then

$E(X) = \lambda$, $Var(X) = \lambda$

Note:

The Poisson probability distribution provides a close approximation to the binomial probability distribution when n is large and p is quite small or quite large with $\lambda = np$.

$$P(X = x) = \frac{(np)^x e^{-(np)}}{x!}, \quad x = 0,1,2,.....$$

$$Where \ \lambda = np = the \ average \ number.$$

Usually we use this approximation if $np \le 5$. In other words, if $n > 20$ and $np \le 5$ [or $n(1-p) \le 5$], then we may use Poisson distribution as an approximation to binomial distribution.

Example:

1. Find the binomial probability P(X=3) by using the Poisson distribution if $p = 0.01$ and $n = 200$

Solution:

$$U\sin g\ Poisson\ , \lambda = np = 0.01*200 = 2$$

$$\Rightarrow P(X=3) = \frac{2^3 e^{-2}}{3!} = 0.1804$$

$$U\sin g\ \ Binomial, \ n = 200, \ \ p = 0.01$$

$$\Rightarrow P(X=3) = \binom{200}{3}(0.01)^3(0.99)^{99} = 0.1814$$

6.4 Common Continuous Probability Distributions

1. **Normal Distribution**

A random variable X is said to have a normal distribution if its probability density function is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \qquad -\infty < x < \infty, \quad -\infty < \mu < \infty, \quad \sigma > 0$$

$$Where\ \ \mu = E(X), \quad \sigma^2 = Var(X)$$

$$\mu \ and \ \ \sigma^2 \ \ are the \ \ Parameters of \ the Normal \ \ Distribution.$$

Properties of Normal Distribution:

1. It is bell shaped and is symmetrical about its mean and it is mesokurtic. The maximum ordinate is at $x = \mu$ and is given by

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}$$

2. It is asymptotic to the axis, i.e., it extends indefinitely in either direction from the mean.
3. It is a continuous distribution.
4. It is a family of curves, i.e., every unique pair of mean and standard deviation defines a different normal distribution. Thus, the normal distribution is completely described by two parameters: mean and standard deviation.
5. Total area under the curve sums to 1, i.e., the area of the distribution on each side of the mean is 0.5. $\Rightarrow \int_{-\infty}^{\infty} f(x)dx = 1$
6. It is unimodal, i.e., values mound up only in the center of the curve.
7. $Mean = Median = \mathrm{mod}\, e = \mu$
8. The probability that a random variable will have a value between any two points is equal to the area under the curve between those points.

Note: To facilitate the use of normal distribution, the following distribution known as the standard normal distribution was derived by using the transformation

$$Z = \frac{X - \mu}{\sigma}$$

$$\Rightarrow f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

Properties of the Standard Normal Distribution:

- Same as a normal distribution, but also...
  - Mean is zero
  - Variance is one
  - Standard Deviation is one

- Areas under the standard normal distribution curve have been tabulated in various ways. The most common ones are the areas between

$$Z = 0 \ and \ a \ positive \ value of \ Z.$$

- Given a normal distributed random variable X with

$$\text{Mean } \mu \ and \ standard \ deviation \ \sigma$$

$$P(a < X < b) = P\left(\frac{a-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{b-\mu}{\sigma}\right)$$

$$\Rightarrow P(a < X < b) = P(\frac{a-\mu}{\sigma} < Z < \frac{b-\mu}{\sigma})$$

Note:

$$\begin{aligned} P(a < X < b) &= P(a \le X < b) \\ &= P(a < X \le b) \\ &= P(a \le X \le b) \end{aligned}$$

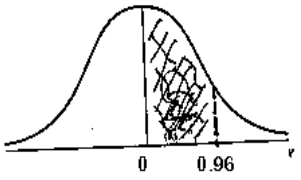
Examples:

1. Find the area under the standard normal distribution which lies

a) Between $Z = 0$ *and* $Z = 0.96$

Solution:

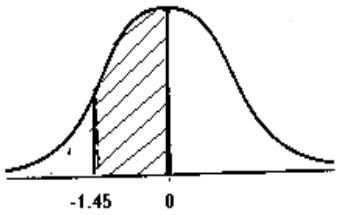
$$Area = P(0 < Z < 0.96) = 0.3315$$



b) Between $Z = -1.45$ *and* $Z = 0$

Solution:

$$\begin{aligned} Area &= P(-1.45 < Z < 0) \\ &= P(0 < Z < 1.45) \\ &= 0.4265 \end{aligned}$$



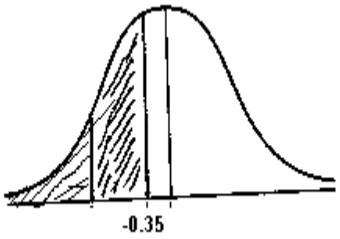
c) To the right of $Z = -0.35$

Solution:

$$\begin{aligned} Area &= P(Z > -0.35) \\ &= P(-0.35 < Z < 0) + P(Z > 0) \\ &= P(0 < Z < 0.35) + P(Z > 0) \\ &= 0.1368 + 0.50 = 0.6368 \end{aligned}$$

d) To the left of $Z = -0.35$

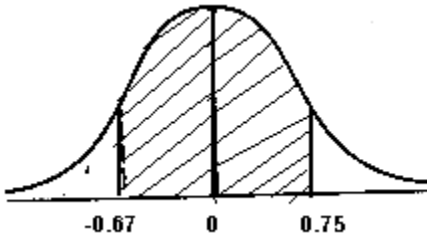
Solution:



$$Area = P(Z < -0.35)$$
$$= 1 - P(Z > -0.35)$$
$$= 1 - 0.6368 = 0.3632$$

e) Between $Z = -0.67$ *and* $Z = 0.75$

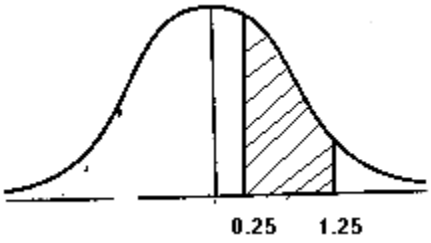
Solution:



$$Area = P(-0.67 < Z < 0.75)$$
$$= P(-0.67 < Z < 0) + P(0 < Z < 0.75)$$
$$= P(0 < Z < 0.67) + P(0 < Z < 0.75)$$
$$= 0.2486 + 0.2734 = 0.5220$$

f) Between $Z = 0.25$ *and* $Z = 1.25$

Solution:

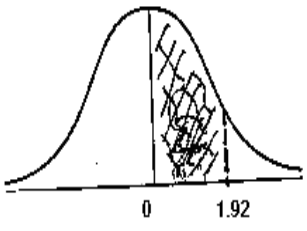


$$Area = P(0.25 < Z < 1.25)$$
$$= P(0 < Z < 1.25) - P(0 < Z < 0.25)$$
$$= 0.3 9 3 4 - 0.0 9 8 7 = 0.2 9 5 7$$

2. Find the value of Z if

a) The normal curve area between 0 and z(positive) is 0.4726

Solution



$P(0 < Z < z) = 0.4726$ *and from table*

$P(0 < Z < 1.92) = 0.4726$

$\Leftrightarrow z = 1.92$ *...uniqueness of Areea.*

b) The area to the left of z is 0.9868

Solution

$$P(Z < z) = 0.9868$$
$$= P(Z < 0) + P(0 < Z < z)$$
$$= 0.50 + P(0 < Z < z)$$
$$\Rightarrow P(0 < Z < z) = 0.9868 - 0.50 = 0.4868$$
$$and \quad from \quad table$$
$$P(0 < Z < 2.2) = 0.4868$$
$$\Leftrightarrow z = 2.2$$

3. A random variable X has a normal distribution with mean 80 and standard deviation 4.8. What is the probability that it will take a value

a) Less than 87.2
b) Greater than 76.4
c) Between 81.2 and 86.0

Solution

$$X\ is\ normal\ \ with\ mean, \mu = 80,\ \ s\tan dard\ \ deviation, \sigma = 4.8$$

a)

$$P(X < 87.2) = P(\frac{X - \mu}{\sigma} < \frac{87.2 - \mu}{\sigma})$$
$$= P(Z < \frac{87.2 - 80}{4.8})$$
$$= P(Z < 1.5)$$
$$= P(Z < 0) + P(0 < Z < 1.5)$$
$$= 0.50 + 0.4332 = \underline{\underline{0.9332}}$$

b)

$$P(X > 76.4) = P(\frac{X-\mu}{\sigma} > \frac{76.4-\mu}{\sigma})$$
$$= P(Z > \frac{76.4-80}{4.8})$$
$$= P(Z > -0.75)$$
$$= P(Z > 0) + P(0 < Z < 0.75)$$
$$= 0.50 + 0.2734 = \underline{\underline{0.7734}}$$

c)

$$P(81.2 < X < 86.0) = P(\frac{81.2-\mu}{\sigma} < \frac{X-\mu}{\sigma} < \frac{86.0-\mu}{\sigma})$$
$$= P(\frac{81.2-80}{4.8} < Z < \frac{86.0-80}{4.8})$$
$$= P(0.25 < Z < 1.25)$$
$$= P(0 < Z < 1.25) - P(0 < Z < 1.25)$$
$$= 0.3934 - 0.0987 = \underline{\underline{0.2957}}$$

4. A normal distribution has mean 62.4.Find its standard deviation if 20.0% of the area under the normal curve lies to the right of 72.9

Solution

$$P(X > 72.9) = 0.2005 \Rightarrow P(\frac{X-\mu}{\sigma} > \frac{72.9-\mu}{\sigma}) = 0.2005$$
$$\Rightarrow P(Z > \frac{72.9-62.4}{\sigma}) = 0.2005$$
$$\Rightarrow P(Z > \frac{10.5}{\sigma}) = 0.2005$$
$$\Rightarrow P(0 < Z < \frac{10.5}{\sigma}) = 0.50 - 0.2005 = 0.2995$$
$$And \quad from\ table \quad P(0 < Z < 0.84) = 0.2995$$
$$\Leftrightarrow \frac{10.5}{\sigma} = 0.84$$
$$\Rightarrow \sigma = \underline{\underline{12.5}}$$

5. A random variable has a normal distribution with $\sigma = 5$.Find its mean if the probability that the random variable will assume a value less than 52.5 is 0.6915.

Solution

$$P(Z < z) = P(Z < \frac{52.5 - \mu}{5}) = 0.6915$$
$$\Rightarrow P(0 < Z < z) = 0.6915 - 0.50 = 0.1915.$$
But from the table
$$\Rightarrow P(0 < Z < 0.5) = 0.1915$$
$$\Leftrightarrow z = \frac{52.5 - \mu}{5} = 0.5$$
$$\Rightarrow \mu = \underline{\underline{50}}$$

6. Of a large group of men, 5% are less than 60 inches in height and 40% are between 60 & 65 inches. Assuming a normal distribution, find the mean and standard deviation of heights.
(Exercise)

2. **The Student's t-Distribution**

- Similar to the normal distribution except that:
  - The population variance is not known so that is estimated from samples
  - Sample size, n, is less than 30
  - Table values are read using n-1 degrees of freedom

3. **Chi-square distribution**

- Not applicable to cases where the observations assume negative values
- Its curve is not symmetrical
- As in the case for t-distribution, n-1 is the parameter of the distribution
- Is used in statistical tests of hypothesis concerning variances, independence of two characteristics and goodness-of-fit.

UNIT-SEVEN: SAMPLING THEORY

7.1. Basic Conceptsof Sampling

Dear learners, can you explain the overall basic concepts of sampling theory?

_____

_____

_____

Before we study the procedures of sampling, it is important to define two important terms in statistics: Population and Sample. In statistical language, **population** is the total elements or items under investigation. **Sample** is a part or subset of this population. For instance, if a researcher is interested to study the performance male and female students in Woliso Campus, all students of the college constitutes the population. Among the students if you select some number of female and male students, this collection which is subset of the population is sample. Another simple example, a medical doctor wants to test the presence of HIV of an individual, all the blood in the body is the population whereas that part of blood used for testing purpose is sample blood. You can also consider a soil scientist who wants to detect the acidity level of a particular soil, all the soil around is the population whereas the part of the soil the scientist must take to the laboratory is the sample soil.

In statistics, the sample taken from the population must approximately represent the characteristics of the population. The possibility of reaching valid conclusions concerning a population, on the basis of a sample is based on two important principles. These are:

a. The law of statistical regularity, and
b. The law of inertia of large numbers.

A) The law of statistical Regularity

This law may be stated as follows: "On an average the sample chosen at random from the universe will have the same composition and characteristics as the universe (population.)". Thus, if there are 400 girls and 600 boys in a college, a random selection of 100 students would yield about 40 girls and about 60 boys. Conversely, if a random selection of 100 students from a college reveals 40 girls and 60 boys. It is reasonable to conclude that, if there are 1,000 students in the college about 400 will be girls and about 600 boys. In the latter case, the result obtained from the study of 100 items will be applied to 1,000 items. This is precisely the method of sampling. But before the results of the sample can be applied to universe (population) two conditions must be met:

- Firstly, the sample should be random, that is every item of in the population has an equal chance of being included in the sample.
- Secondly, it should be sufficiently representative

In statistics, there is a basic principle that the larger the number of items, the more reliable is the results obtained there from. Because, it is possible then to avoid the influence of abnormal items on the average. The larger the size of the sample the more reliable is the result because the sampling error is inversely proportional to the square root of the number of item in the sample. i.e.

$$e=\alpha\frac{1}{\sqrt{n}}$$

Where **e** is the sampling error and **n** is the size of the sample, that is, the number of items included in the sample.

Once it is ensured that the sample selected is representative of the population, it is possible for one to depict fairly and accurately the characteristics and composition of the population by studying only a part (sample) of it. Thus, this law is of much importance as it saves times, energy and money by studying only a part of the population and then applying the results obtained by the sample to the whole population.

B) Inertia of Large Numbers

This principle is an extension to the law of statistical regularity. It only states that: "other things being equal, greater the size of the sample, more accurate the results are likely to be". This is because large groups of data have a higher degree of stability than that possessed by small ones. For example, if a coin is tossed 50 times, head may appear 30 times and the tail 20 times. But if the coin is tossed 1,000 times, we may get 500 heads and 500 tails. This is so because when the number is very large, then some item move in one direction, and others move in the opposite direction, thus canceling out each other.

7.2. Reasons of Sampling

Dear learners, can you describe the basic reasons of sampling for statistical investigation?

_____

_____

_____

It is not always possible to examine every member of population (also called the **Universe**) in order to draw conclusions. Hence, there is a needed to use a technique called sampling. Sampling will allow a researcher to find out something about the larger population. **Sample** is defined as subset of units selected from large set of subset of units. The subset provides data for used in estimating the characteristics of the large set. In other words, we use measurements of the characteristics sample for forecasting similar characteristics about larger samples or population. The sample data are used to predict how a population will act or react under the same conditions in some future situation or event.

You resort to sampling, as it is cost, time and labor efficient. It might also result in height level of accuracy. Further, in some cases, it is impossible to identify every member of s population. The following might explain sampling manner. When you cook rice in a pot, you check only a few rice to find out whether the entire rice in the pot has been cooked to the desired extent. A grain merchant might judge the quality of grain in a shipment by examining only a few of the bags.

Sampling is unnecessary if the population to be surveyed is small. For large population, researchers must select samples in unbiased ways to ensure that no group or stratum is systematically over or under represented. A representative sample is one that has every

major attribute of the large population present in bout the same proportion. How can the researcher know whether the sample is representative of a particular population's characteristics or about the opinions of its members? They never know for sure that their sample exactly mirrors the population. If they apply the principle of probability theory, researchers can estimate their sample's accuracy and establish a certain level of confidence in their estimate.

In most cases, researchers will not be in a position to collect information from every member of their selected research population. Either the group is too large, too dispersed, or costs and time restrictions will be prohibitive. The alternative is to gather information from some members of that population who are selected in ways designed to ensure that their responses and characteristics are representative of the whole.

In summary, the following are some of the merits of making use of sampling technique.

- Sampling is best when results are required urgently.
- Sometimesit is impossible to be analyze all elements of the population. For example in testing explosives, sample can be tested to find out the strength of explosive. Another case is to test the effectiveness of drug, we can only take sample of people to find out the effectiveness of the drug on the disease under study.
- Sample method requires less time, money, and manpower as compared to analyzing all elements of the population.
- In the case of sampling, we are in a position to get much more accurate information than is possible by census method:
    - ✓ Detailed information can be obtained from a small group of respondents,
    - ✓ Qualified persons or investigators can be appointed and intensive training given,
    - ✓ Relatively limited data can be handled much more easily,
    - ✓ Follow-up is easy in case of poor-response,
    - ✓ Error that occurred in recording or collecting data (Non-sampling errors) will be minimized as the data collected and processed is relatively small.

However, the following are some of the demerits of sampling method:

- The results obtained may be false, inaccurate and misleading as the sample might not have been drawn properly,
- Chances of sampling errors- errors occurred in taking samples from a population- are great.
- When the population is small, sampling is not useful

7.3. Procedures of Sampling

Dear learners, can you elaborate the procedures to be followed in sample selection?

_____

_____

_____

A common goal of survey research is to collect data representative of a population. The researcher uses information gathered from the survey to generalize findings from a drawn sample back to a population, within the limits of random error.

Thus, samplingis the process by which a relatively small number of individuals or events is selected and studied to find out something about the entire population from which it was selected**.**It is believed that if this sample is chosen carefully using the correct procedure, it is then possible to generalize the results to the whole of the research population.

The following are some of the consideration while developing a sampling design.

Step 1:Define the Population/ Type of Universe:

The first step in developing any sample design is to clearly define the total set of objects in which the researcher is interested. Thus you can define population as a collection of people, animals, plants or things on which you may collect data. It is the entire group of interest which you wish to describe or about which you wish to draw conclusions. It is impractical for an investigator to completely enumerate the whole population for any statistical investigation. It can be finite or infinite.

For example, if you want to have an idea about the average monthly income of people residing in Ethiopia, you will have to enumerate all the earning of individuals in the country, which is rather a very difficult task. Also, when population is large infinite or if units are destroyed during investigation, it is not possible to enumerate or investigate the whole population. But even if population is finite 100% inspection is not possible because of factors such as, time, money and administrative convenience.

If you want study about the Automobile brand preference of car buyers living in Addis Ababa, your population will be all people in Addis Ababa who can afford buy automobile. If you want to know the academic achievement of female students in Woliso Campus, your population may be all female students of Woliso Campus.

Step 2:Specify Sampling Unit

The sampling unit is the **basic unit containing the population** to be sampled. Sampling unit may be a **geographical one**, such as Region, Zone, District, Woreda and Kebele. It may be a **construction unit**, such as house, flat and buildings. It may be a **social unit**, such as family, club, school and community.

Step 3:Specify Sampling Frame

It is the specific set of units from which the sample is actually drawn. This is also known as the **source list**. It contains the **name of all items of a universe** in case of a **finite universe**. The sample has to be drawn from that list. In case a source list/sampling frame is not available for a finite universe, the researcher has to prepare it. Such a list should be **comprehensive**, **correct**, **reliable** and **appropriate**. The sampling frame has to be a **representative of the population**.

- A map, a telephone directory, a list of shops in Mercato that come under Value Added Tax (VAT) are Example of Sampling Frame.

Therefore, sampling frame tells you that **total size of the population** and **helps you decide** the **number of samples to taken** and the **type of sampling to be used**.

Step 4:Determine Sample Size

This is related to the number of items to be selected from the universe to constitute a sample. In this step, the number of elements or units of the population to be sampled will be decided, the size should be optimal and it can neither be expressively large nor too small.

A) Sample size criteria

In addition to the purpose of the study and **population size**, four criteria usually will need to be specified to determine the appropriate sample size: the **level of precision**, the **level of confidence** or risk, the **degree of variability** in the attributes being measured and **response rate**.

1) Level of precision

The *level of precision*, sometimes called *sampling error*, is the margin in which the true value of the population is estimated to be. This range is often expressed in percentage points of a plus and minus figure that represents how accurately the answers given by the given sample size correlate to the answers given by the entire population.

Example 1: If a researcher finds that 60% of the farmers in the sample have adopted a recommended practice with a precision rate of **+/- 5%,** then he/she can conclude that between 55% and 65% of farmers in the population have adopted the practice.

Example 2: If 65% of the respondents in the sample are very much satisfied 'with the service of Ethiopian Airlines and the margin of error is **+/- 4%,** then the sample result provides an estimate the sample result provides an estimate that between 61% and 69% of the target population is very much satisfied with the service.

- **NB**: 5% is the most commonly used margin of error, but we may want anywhere from 1% to 10% for a margin of error depending on our survey. Increasing the margin of error above 10% is not recommended.

2) Level of confidence

It means how sure you can be that a particular sample's estimates fall within a specified range of a statistic. Most public administrator will accept a confidences level of 95% or 99%. It means that if you were to draw 100 separate samples of a population the sample estimates of the population parameter would fall within a designated range of acceptable error in 95(or 99) of the 100 samples.

Theconfidence or risk level is based on ideas encompassed under the Central Limit Theorem. The key idea in the Central Limit Theorem is that when a population is repeatedly sampled, the average value of the attribute obtained by those samples is equal to the true population value.

Example 1: If sample is a certain size, you can say that you are 95% sure that between 61% and 69% of the population are 'very much satisfied' with the service of Ethiopian Air lines. There is a 5% chance that the actual population parameter falls outside this range. Most of the researchers are willing to accept these odds. In other words, this means that if a 95% confidence level is selected, 95 out of 100 samples will have the true population value within the range of precision specified earlier.

- **NB**: 95% is the most commonly used confidence level but you can use a range between 90% to 99% confidence level depending on your survey. Decreasing the confidence level below 90% is not recommended.

3) Degree of variability

It refers to the distribution of attributes in the population. The more heterogeneous a population, the larger the sample size required to obtain a given level of precision. The less variable (more homogenous) a population, the smaller the sample size. Note that a proportion of 50% indicates a greater level of variability than either 20% or 80%. This is because 20% and 80% indicate that a larger majority do nor or do, respectively, have the attribute of interest.

4) Response rate

It refers to the percentage of people who do actually fill out the survey that they receive. It varies widely depending on a number of factors such as the relationship with your target audience, survey length and complexity, incentives and topic of the study. Estimating the response rate will help you determine the total number of surveys you will need to send out to obtain the required number of completed surveys.

B) Strategies to Determine Sample Size

1) Using a Census for Small Population

One approach which uses the entire population as the sample. Although cost consideration make this impossible for large populations, a census is attractive for small populations (e.g., 200 or less). A census eliminates sampling error and provides data on all the individuals in the population at desired level of precision.

2) Using/Imitating a Sample Size of a Similar Study

Another approach is to use the same sample size as those of studies similar to the one you plan. Without reviewing the procedures employed in these studies you may run the risk of operating errors that mere made in determining the sample size for another study. However, a review of literature in your discipline can provide guidance about "typical" sample sizes that are used.

3) Using Formulas

Although tables can provide a useful guide for determining the sample size, you may need to calculate the necessary sample size for a different combination of levels of precision, confidence and variability. The third approach to determine sample size is the application of several formulas.

Formula One: Israel, 1991

Case 1: For **populations that are too large**, Cochran developed the following equation to yield a representative sample for proportions.

- $\mathbf{n_o = \frac{Z^2pq}{e^2}}$

where $\mathbf{n_o}$ is the sample size, $\mathbf{Z^2}$ is the abscissa of the normal curve that is the computed **Z-score** (based on the confidence level), **e** is the desired level of precision (margin of error), **p** is the estimated proportion of an attribute that is present in the population (the percentage of the sample who will respond a given way) and **q** is **1-p**.

- The value of **Z** is found in statistical tables which contain the area under the **normal (Z) curve**.
- We may use anywhere from 1% to 10% for a margin of error depending on our survey.

To illustrate, suppose we wish to evaluate a state-wide extension program in which farmers were encouraged to adopt a new practice. Assume there is a large population but that we do not know the variability in the proportion that will adopt the practice; therefore, assume p=50%=0.5 (maximum variability). Furthermore, suppose, we desire a 95% confidence level and +/- 5% precision level/margin of error. At 95% confidence level, the computed standard (Z score) value from the normal distribution table is 1.96.

The resulting sample size is demonstrated as:

- $\mathbf{n_o = \frac{Z^2pq}{e^2} = \frac{1.96^2(0.5)(0.5)}{0.05^2}}$ **= 385 Farmers**

Case 2: if the **population is small** then the sample size can be reduced slightly. This is because a given sample size provides proportionately more information for a small population than for a large population. The sample size (n_o) can be adjusted using:

- $\mathbf{n = \frac{n_o\,(N)}{N+(n_o-1)}}$

- where **n** is the **sample size** and **N** is the **population size**

Example: Suppose our evaluation of farmers' adoption of the new practice only affected 2, 000 farmers. The sample size that would now be necessary is shown as:

- **n =** $\frac{\mathbf{n_o\ (N)}}{\mathbf{N+(n_o-1)}} = \frac{\mathbf{385\ (2000)}}{\mathbf{2000+(385-1)}}$ **= 323 Farmers**

As you can see this adjustment called (**Finite Population Correction-FPC**) can substantially reduce the necessary sample size for small populations.

Formula Two: Yemane, 1967

Yemane provides a simplified formula to calculate sample sizes. This formula is used to calculate the sample sizes in the table below. A 95% confidence level and P=0.5 are assumed for the formula.

- **n =** $\frac{N}{1+N(e^2)}$**,** where **n** - sample size, **N** - population size &**e** - margin of error

Example: suppose this formula is applied to the above sample, we get:

- **n =** $\frac{N}{1+N(e^2)} = \frac{2000}{1+2000(0.05^2)}$ **= 333 Farmers**

4) Using Published Tables

A fourth way to determine sample size to rely on published tables, which provide the sample size for a given set of criteria.

N	Sample size (n) when e =					N	Sample size (n) when e =			
	3%	5%	7%	10%			3%	5%	7%	10%
500	a	**222**	145	83		**100**	a	**81**	67	51
600	a	**240**	152	86		**125**	a	**96**	78	56
700	a	**255**	158	88		**150**	a	**110**	86	61
800	a	**255**	163	89		**175**	a	**122**	94	64
900	a	**267**	166	90		**200**	a	**134**	101	67
1,000	a	**286**	169	91		**225**	a	**144**	107	70
2,000	714	**333**	185	95		**250**	a	**154**	112	72
3,000	811	**353**	191	97		**275**	a	**163**	117	74
4,000	870	**364**	194	98		**300**	a	**172**	121	76
5,000	909	**370**	196	98		**325**	a	**180**	125	77
6,000	938	**375**	197	98		**350**	a	**187**	129	78
7,000	959	**378**	198	99		**375**	a	**194**	132	80
8,000	976	**381**	199	99		**400**	a	**201**	135	81
9,000	989	**383**	200	99		**425**	a	**207**	138	82
10,000	1,000	**385**	200	99		**450**	a	**212**	140	82
15,000	1,034	**390**	201	99						
20,000	1,053	**392**	204	100						

25,000	1,064	**394**	204	100						
50,000	1,087	**397**	204	100						
100,000	1.099	**398**	204	100						
>100,000	1,111	**400**	204	100						

➢ **NB: a** = assumption of normal population is poor. The entire population should be sampled.

7.4. Techniques of Sampling

Dear learners, can you explain the two major techniques of sample selection from a given target population?

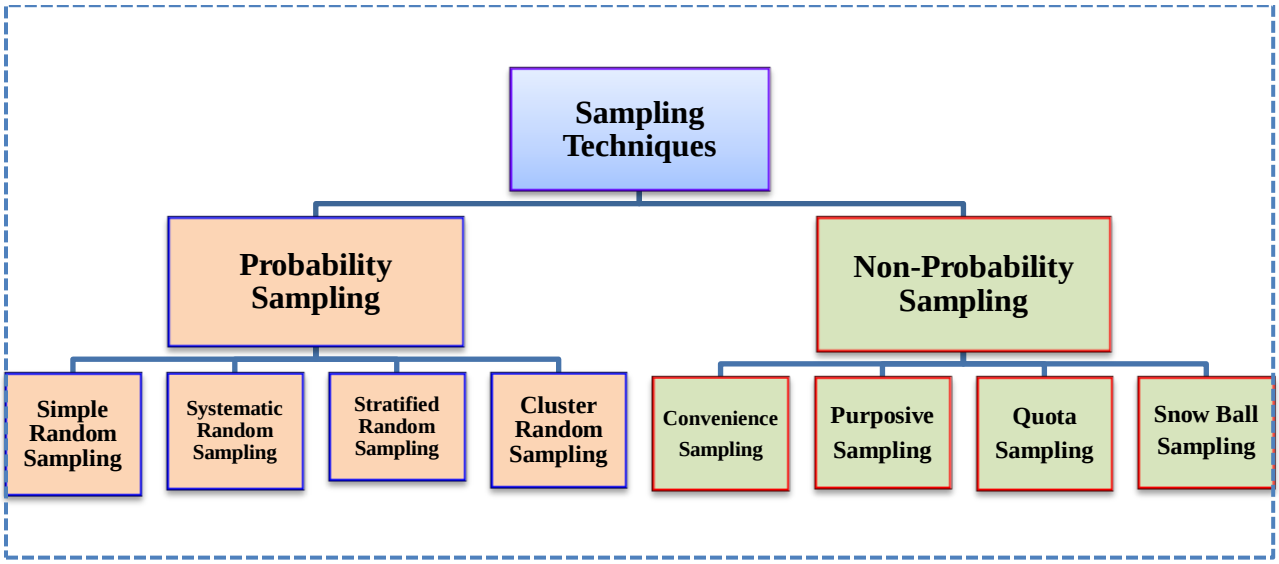
_____

_____

_____

Sampling techniques are the different techniques of collecting data (information) from a portion of a population. These are main divisions of sampling methods in which we take sample from a population. The major sampling techniques, as indicated in the following diagram, may be grouped in to.

- probability sampling (Random Sampling)
- non-probability sampling (Non-Random Sampling)



A) Probability Random Sampling

It is based on chance selection procedure i.e., every element of the population has a known non-zero probability of selection. This eliminates the bias inherent in the non-probability sampling procedures because probability sampling process is random.

Randomness refers to a procedure the outcome of which cannot be predicted because it is dependent on chance.

- The selection of the sample is based on the theory of probability is also known as **random selection** and sometimes probability sampling is also known as random sampling.

In other words, probability sample is based on the concept of random selection by which each unit in the population has some chance of being in the sample; that **chance is greater than zero** and **can be calculated**. Moreover, **probability samples are preferred** if **accurate inference** are to be made of population characteristics on the basis of sample statistics.

There are different techniques of probability sampling techniques. The **common techniques** are indicated as:

1) Simple random sampling
2) Systematic random sampling
3) Stratified random sampling, and
4) Cluster random sampling

1) Simple Random Sampling

It is a probability sampling in which each element in the population has an equal chance of being included in the sample. The selection of each unit is independent of the selection of any other unit i.e. the selection of a member of the population for the sample should not increase or decrease the probability that any other member of the sample. The sampling process is simple because it requires only one stage of sample selection.

There are two common methods of constructing random sample. These are:

(a) Lottery Method

- Generally you make one card/slip corresponding to one unit of population by writing on it the number assigned to that particular unit of population for sampling purposes.

- The card/slip are shuffled a number of times and then a card is drawn at random from them until the sample size.
- This is one of the most reliable methods of selecting a random sample.

(b) Random Number Table Method

The explained method of lottery is time consuming and cumbersome to used if population is very large. Therefore, the most practical and inexpensive method of selecting a Random Number Table which has been constructed that each of the digits 0,1,2,3,4,5,6,7,8,9 appear with approximately the same frequency and independently of each other.

If you have to select a simple random sample from a **population of sized N(d"99)** then the numbers can be combined **two by two** to give pair from **00 to 99**. Similarly, if **Nd"999** or and so on, then combining the digits **three by three** (or **four by four** and so on), you get numbers from **000 to 999** or (**0000 to 9999**) and so on since each of the digits 0,1,2,3,4,5,6,7,8,9, appear with approximately the same frequency and independently of each other, so does each of the pair 00 to 99 or triplet from 000 to 999 or quadruplets 0000 to 9999 and so on.

Thus, the method of drawing the random sample consists in the following steps:

i. **Identify** the **N unit** in the population with the **number** from **1** to **N**;
ii. **Select** at **random** any page of the random numbers in any **row** or **column** or **diagonal** at random; and
iii. The **population units** corresponding to the **number of unit selected** in step (ii) **comprise the random sample**.

Example: The following table shows the partial list numbers on a random number table

.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....
.....	.....	.....	.....	.....	**16376**	**39400**	**53537**	**71341**
.....	.....	.....	.....	.....	**91782**	**60468**	**81305**	.....
.....	.....	.....	.....	.....	**53498**	**18672**	.....	.....

.....	.....	.....	.....	.....	**31016**	**71194**	.....	.....
.....	.....	.....	.....	.....	**20922**	.....	.....	.....
.....	.....	.....	.....	.....	**18103**	.....	.....	.....
.....	.....	.....	.....	.....	**59533**	.....	.....	.....

Let's look at the following steps to select a sample using a random number table.

i. **Assign numbers** for members the study population from 1 to N.

ii. Use a random number table and select any number from anywhere on the number table starting from any raw or column on number table & move to any direction that you want.

- ✓ For example, you can take the number coinciding to the **4th row** and the **6th column**, which is **16376** which is **your starting point**.

iii. The number of digits you will read depends on the total number of the study population=N.

- ✓ Let's say if it is **N=90**; you will **read two digits**, if **N=150**, you will **read three digits** and so on.

iv. The number of digits you will read will be selected randomly.

- ✓ For example; if **N=150** you start from with the **first three digits** (163) or **middle three digits** (637) or the **last three digits** (376).

v. Then begin to read in any direction to the right or to the left or up or down

vi. Ignore those numbers that are greater than N and those which may appear two times.

vii. Stop counting after you get the required number of sample subjects (n).

- ✓ For example, **N=150**(total population); **n=20** (sample size);
- ✓ Let's take **16376 as a starting point** and consider the **last three digits (376),** and read **downward**. The **first number** is **376**; which is **greater than 150**. So, leave it. Also **782,498,533** are **greater than N**.
- ✓ So, your subject holding **016,103** will become your **members for sample selection**. You will **proceed to the next sample selection** activities **until you will get 20 subjects**.

Advantages of Simple Random Sampling

- ✓ Since sample units are selected at random providing equal chance to each and every unit of population to be selected, the element of subjectivity or personal bias is completely eliminated.
- ✓ You can ascertain the efficient of the estimates of the parameters by considering the sampling distribution of the statistic (estimates).
- ✓ Easy to analyze data and compute error for homogenous distribution of the population.

Limitations of Simple Random Sampling

- ✓ The selection of simple random sample requires an up-to- date frame of population from which samples are to be draw although, it is impossible to have knowledge about each and every unit of then population happens to be very large. This restricts the use of simple random sample.
- ✓ A simple random sample may result in the selection of the sampling units, which are widely spread geographically; and in such a case the administrative cost of collecting the data may be high in term of time and money.
- ✓ For a given precision, simple random sample, usually, requires large sample size compared to stratified random sampling which will be discussed next,

2) Systematic Random Sampling

A sampling procedure in which an initial starting point is selected by a random process, and then every n^{th} number on the list is selected. Because the 1^{st} element selected is a random choice, a systematic sample is usually assumed to have the properties of a simple random sample. This assumption is especially applicable when the list of elements in the population is a random ordering of the elements.

In general, Systematic Random Sampling uses the following guiding procedures:

- ✓ **Step 1-**Arranging elements of population in some systematic order based on the objective of the statistical investigation,
- ✓ **Step 2-**Divide the number of population (**N**) to the number of sample elements (**n**) to be selected. The result is called Skip interval.

- **k =** $\frac{N}{n}$**,** where **k** is **skip interval**

✓ **Step 3**- Select an initial starting point 'S_1'from '**1^{st} to k^{th},** elements of the arranged population using simple random sampling method, where **S_1** is considered as the 1^{st} element of the sample,and its value, in fact, determines the whole sample.

✓ **Step 4**-Every k^{th}element of the population on the list will be selected until all sample elements are achieved. In other words, to select the remaining elements of the sample, we can apply "**Arithmetic Sequence**."

Advantagesof Systematic Random Sampling

- It is operationally more convenient than simple or stratified random sampling techniques as it saves you time and work involved

Disadvantages of Systematic Random Sampling

- If sampling interval is related to a periodic ordering of the population, it may introduce increased variability.

3) Stratified Random Sampling

In stratified random sampling, the first step is to divide or classify the population into groupscalled strata.The basis for forming strata, such as departments, sex, location, age or institution and so on is at the discretion of the designer of the sample. The classification has to be done so that every member of the population is found in one and only one stratum i.e. **classification** should be **exhaustive** and **mutually exclusive.**After the strata are formed, separate random sample are drawn from each stratum using simple random sampling. The researcher determines the number of each unit selected from each stratum. Hence, stratified sample provide for greater accuracy than simple random sample of the same size.

Stratified sampling can be a very efficient way to reduce sampling error and increase the representativeness of a sample. A researcher can reduce sampling error by increasing the sample size or by reducing population variability. Stratified sampling reduces population variability and lowers sampling error in estimating a population parameter by ensuring that different groups are adequately represented in the same.It ensures that differences

across the strata on some dependent variable of interest are accounted for and are not free to vary in the sample. A smaller stratified variable is related to the dependent variable of interest.

There are two type of stratified random sampling know as:

(a) Proportionate Stratified Sampling; and (b) Disproportionate Stratified Sampling.

i) Proportionate Stratified Sampling

Members of a population are classified into strata. Number of units selected from each stratum is directly proportional to the size of the population in that stratum. For example, the productivity of field crops can be stratified based on economic use, such as cereals, pulses, beverages, vegetables, fruits, etc. An equal percentage of number from each stratum would be drawn for the sample would consist of five strata, each equal in size to the stratum's proportion to the total population.

ii) Disproportionate Stratified Sampling

A large percentage is taken from some strata than other. Researcher selects a larger percentage from group likely to be underrepresented in the population. For example, there might be very few crops in the stratum of beverages. In such a case, the researcher might take a larger percentage from this stratum so that adequate number of crops represents this stratum. Simple random sample or proportionate sample may have too few members with the characteristic to allow full analysis. Number of units included in each stratum of the sample will be large enough to allow for separate analysis for each individual stratum. Since some strata have been sampled at a higher percentage than other (they have been over-sampled) a weighing procedure must be applied when they are combined to form one sample.

Advantages of Stratified Sampling

- It assures representation of all groups in a sample.
- Characteristics of each stratum can be estimated and comparisons made.

- Further it reduces variability for same sample size.

Disadvantages of Stratified Sampling

- It requires accurate information on proportion in each stratum.
- If stratified lists are not already available they can be costly to prepare.

4) Cluster Random Sampling

Cluster sampling also called **multistage sampling** or **area sampling** is useful when a list of the entire target population (**sampling frame**) is **unavailable** or **impractical to compile**. To this end, drawing a simple random sample or stratified sample of a population, such as, regional residents could be very difficult. This method can reduce the cost of data collection for large dispersed populations'.

In cluster sampling, the elements in the population are **first divided into separate groups** called **clusters**, such as zones, districts,woredas and kebeles. Each element of the population belongs to one and only one cluster. All elements within each sampled cluster form the sample. If all the units within selected cluster are included in the sample, the design is called **clued sampling**. Cluster sampling tends to provide the best results when the elements within the clusters are not alike.

Unlike the stratified sampling, which draw cases from every stratum in a population, cases are selected only from certain clusters. In **single–state cluster sampling**, there is no sampling error within a cluster because everyone in the cluster is interviewed. Sampling error occurs only between clutters. Therefore it is important that the researcher select a sample of clusters proportionate to their size in the population. A stratification by geographic region also help to ensure that the opinion of residents any particular sector are represented.

One of the primary applications of cluster sampling is area sampling, where clusters are city blocks or other well defined areas. Cluster sampling generally requires a larger total sample size than either simple random sampling or stratified random sampling. However, it can result in cost savings because of the fact that when an interviewer is sent to a

sampled cluster, many observations can be obtained in a relatively short time. Hence, a larger sample size may be obtainable with a significantly lower total cost. The value of cluster sampling depends on how representative each cluster is of the entire population. If all clusters are alike in this regard, sampling a small number of clusters will provide good estimates of the population parameters.

B) Non-Probability Random Sampling

The selection for elements in non-probability sampling is quite arbitrary, as researchers rely heavily on personal judgment, beliefs, experience and expertise.In this method, some members of the eligible target population have a chance of being chosen and others do not. It is easier, quicker and cheaper to carry out than probability designs. At the same time, it is impossible to determine, scientifically, how represent the sample is of some large population. They are not desirable for use in gathering data to be analyzed by the methods of inferential statistics.

The following are some of the type of non-probability sampling design to be discussed in the subsequent section.

1) Convenience Sampling
2) Judgment or purposive sampling
3) Quota sampling
4) Snow Ball Sampling

Non- probability sampling methods have limitation since they do not employ random selection of cases. It is impossible to know how accurately the result of such survey represents the larger population. A researcher cannot compute a confidence level or an error margin for non-probability sample. The classic example of a non- probability sample is a questionnaire published in a news paper or magazine or in the website not everyone in the target population reads or subscribes to his paper of magazine or has access to internet. This type of survey violates the major principle of probability theory.

1) Convenience Sampling

This sampling technique involves sampling on the basis of availability of units, nearness, and/or willingness to participate.Example: Interview conducted in convenient location, such as meeting places or celebration places. It is unwise to use this sampling to making inferences about general populations. If the purpose is to identify issues of potential concern to a large population, convenience sampling may be both economical and appropriate the subjects are selected non- randomly simple because they are viable. For example, market researchers, often go to shopping malls and try to interview anyone who come through the doors.

Convenience samples have the advantage of relatively easy sample selection and data collection; however, it is impossible to evaluate the "goodness" of the sample in terms of its representativeness of the population. A convenience sample may provide good results or it may not; not statistically justified procedure allows a probability analysis and inference about the quality of the sample results.

Advantages

- Very low cost, extensively used.
- No need for list of population, i.e., it is independent of the size of the population.

Disadvantages

- Variability and bias of estimates cannot be measured or controlled

2) Purposive Sampling / Judgment or Expert Choice Sampling

Researchers' judgment that the units somehow represent the population is the main criterion in the main criteria of the judgment or expert choice sampling. Usually, the person who makes the selections is an expert in the area concerned unit for this type of sample are selected on the basis of known characteristics that seem to represent the population. It is assumed that the unit selected will represent the population on unknown

characteristics as well. This type of sampling lacks one or more of the conditions of the probability sampling. For instance; it may not have the randomness of a sample.

It is non-probability sampling technique in which an experienced individual (expert) selects the sample based upon his or her judgment about some appropriate characteristic required of the sample members. Usually, the person who makes the selections is an expert in the area concerned. However, the quality of the sample results depends on the judgment of the person selecting the sample.

Advantages

- It is useful for certain types of forecasting like sample guaranteed to meet a specific objective.
- It has moderate cost and average use.

Disadvantages

- It introduces bias due to experts' beliefs and it may make sample unrepresentative.
- because elements in the population don't have some chance to be included in the sample

3) Quota Sampling

In quota sampling, various segments of a population have the same percentage of representation in the sample as they have in the population. However, the elements in the sample are not selected randomly, rather based on judgment. Quota sampling is the most widely-used form of non-probability sampling. In this case some strata or grouping population will be considered in sample selection. The grouping variables' may be age, sex, race, area, etc. The allocation of sample units will be simply determined as a quota. For example, interviewers are given targets or quotas to achieve such as '10 women aged 40-50' or '20 middle class men aged over 35'.

It is non-probability sampling in which the researcher classifies population by pertinent properties, determines desired proportion of sample from each class and quotas for each. The first step of choosing strata on the basis of existing information is the same for both

stratified and quota sampling. However, the processes of selecting sampling units (elements) within the stratum differ substantially. In stratified sampling, a sub sample is drawn using simple random sample within each stratum. This is not true with quota sampling. The purpose of quota sampling is to ensure that the various subgroups in a population are represented on pertinent sample characteristics to the exact extent that the investigators desire.

Advantages

- It introduces some stratification of population and requires no list of population.
- It has moderate cost and it is used very extensively.
- One can finish data collection in a very short period of time.

Disadvantages

- It introduces bias in researcher's classification of subjects.
- Further non-random selection within classes means error from population cannot be estimated.

4) Snow Ball Sampling

In snowball sampling, you begin by identifying someone who meets the criteria for inclusion in your study. You then ask them to recommend others who they may know who also meet the criteria. Although this method would hardly lead to representative samples, there are times when it may be the best method available. Snowball sampling is especially useful when you are trying to reach populations that are inaccessible or hard to find.

- For instance, if you are studying the homeless, you are not likely to be able to find good lists of homeless people within a specific geographical area. However, if you go to that area and identify one or two, you may find that they know very well who the other homeless people in their vicinity are and how you can find them.

✍ Self-Assessment Question-Unit 6 (Sampling Theory)

Attempt all the following questions. Then check your answer against the answer key provided at the end of the module.

Part I: Matching

	"X"		"Y"
1	It is commonly called multistage sampling or area sampling	**A**	Simple random sampling
2	It assures representation of all groups in a sample using random probability selection	**B**	Convenience sampling
3	Sample selection in a nearby area, meeting places, market areas or celebration places.	**C**	Systematic random sampling
4	It is also called Quasi Random Sampling as selection of one unit affects the probability of selection of other units.	**D**	Purposive sampling
5	Sample selection is based upon his/her judgment about some appropriate characteristic required of the sample members by an experienced individual (expert)	**E**	Stratified random sampling
6	It expresses how much the characteristics of a sample statistics differ from the parameters of the population.	**F**	Quota sampling
7	Each element in the population has an equal chance of being included in the sample.	**G**	Cluster random sampling
8	Errors that occur for other reasons, such as errors made during collection, recording, and tabulation of data	**H**	Snow ball sampling
9	It is especially useful when you are trying to reach populations that are inaccessible or hard to find.	**I**	Sampling error
10	Ensure that the various subgroups in a population are	**J**	Non-sampling error

	represented on pertinent sample characteristics to the exact extent that the investigators desire.		

References

1. Bluman, A. G. (1995). Elementary Statistics: A step by step approach (2nd Edition).W.MC. Brown Communication, Inc.
2. Coolidge, F.L. (2006). A Gentle Introduction (2nd Edition)
3. David, S.M., Mc Cabe, P. and Craig, B. (2008). Introduction to the practice of Statistics (6th Edition). W.H. Freeman
4. Eshetu W. (2000). Introduction to statistics. Addis Ababa University Press