AMBO UNIVERSITY WOLISO CAMPUS

SCHOOL OF BUSINESS AND ECONOMICS

DEPARTMENT OF AGRICULTURAL ECONOMICS

INTRODUCTION TO STATISTICS

Course Code: (ABVM2101)

(3/5 Cr.Hrs/ECTS)

Module Writer:

Getahun G. Woldemariam (MSc)

May, 2020

Woliso, Ethiopia

TABLE OF CONTENTS

MODULE OBJECTIVES

The learning task was designed to equip students with the ability to

- ☞ **Identify the importance and application areas of statistics in their field of study;**
- ☞ **Interpret statistical information, reports, charts and figures;**
- ☞ **Choose appropriate sampling methods and procedures;**
- ☞ **Explain the basic concepts of probability distributions and their application;**
- ☞ **Use estimation and testing methods for predication and generalization purposes.**
- ☞ **In addition, the learning task attempts to enable students to describe data collection tools and procedures.**

MAJOR COMPONENTS

1. Lectures

NO.	Title	Hours
L1.1	Introductory concepts in statistics	4
L1.2	Measures of central tendency	5
L1.3	Measures of dispersion	5
L1.4	Probability theories	8
L1.5	Concepts of sampling and their applications	5
L1.6	Estimation and hypothesis testing	8
L1.7	Correlation and simple linear regression	5

2. Problem based learning tasks

Number	Case Description
PBL1.1	Students will be provided with socio-economic data and asked to compute various measures such as frequency distribution, measures of central tendency, measures of dispersion, and measures of shape of distribution

3. Individual Studies

NO	Title of Book, Article or website; or Reader	Hrs
S1.1	Agrawal B.L. 1996. Basic statistics, new age international pub. Ltd. New Delhi	7
S1.2	Frank H. and Althoen S.C. 1994. Statistics: concepts, and application. Cambridge university press, UK	7
S1.3	Hooda R.P, 2001. Statistics for business and economics. 2nd , New York	7
S1.4	Johnson , R.A, and Bhata K.G. 1992, statistics principles and methods. New York	7
S1.5	Wayne, W. 1995. Biostatistics : a foundation for analysis in health. 6th ed. New York	7
S1.6	Students read their handouts, notes and any other materials they find helpful to fulfill the objectives of the educational unit	20

4. Practical Activities:

NO	Title Practicals (guided)	Hou
P1.1	With the help of appropriate software, students will be asked to compute correlation and various test of hypothesis based on fictitious socio-economic data	3
Demonstrations		
D1.1	Students follow instructor's demonstration of software package application and exercise to master the application	4
Routine training (independent)		
R1.1	Experts from statistical offices and other institutions that are known to use data processing activities will be invited to train students and share their practical experiences.	3

L task	Total hrs LT	Hrs for different educational activities within the task (LT)						
		L	P	T	S/WS	PA	A/PoA	IS
LT 1	135	40	10	8	7	8	7	55

L1.1 INTRODUCTORY CONCEPTS IN STATISTICS

1.1.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 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.

1.2.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.

1.3.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.)

1.4.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

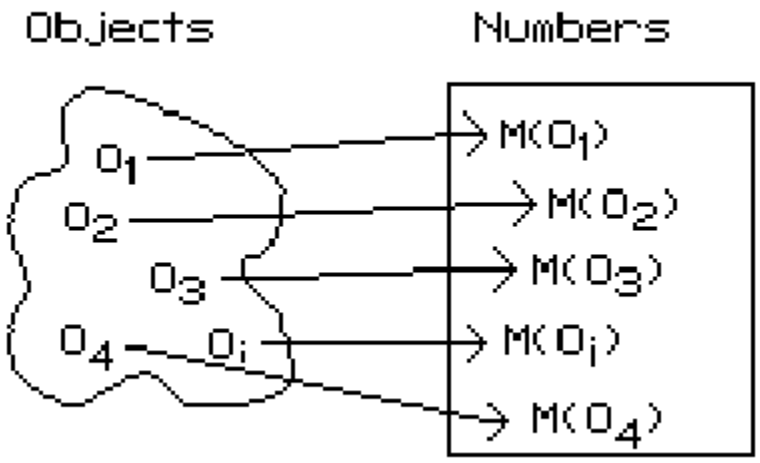
As a science statistics has its own limitations. The following are some of the limitations:

- Deals with only 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.

1.5.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.

1.6.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 oF

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)

- Your checking account balance as a measure of the amount of money you have in that account.
- Your score on the first statistics test as a measure of your knowledge of statistics.
- Your score on an individual intelligence test as a measure of your intelligence.
- The distance around your forehead measured with a tape measure as a measure of your intelligence.
- 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.
- Times for swimmers to complete a 50-meter race
- Months of the year Meskerm, Tikimit…
- Socioeconomic status of a family when classified as low, middle and upper classes.
- Blood type of individuals, A, B, AB and O.
- Regions numbers of Ethiopia (1, 2, 3 etc.)
- The number of students in a college;
- The net wages of a group of workers;
- the height of the men in the same town;

L1. 2. METHODS 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 is 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.

2.1.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 (1)	Tally (2)	Frequency (3)	Percent (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 the entire steps one can construct the following frequency distribution.

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

2.2.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.

2.2.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 below:** 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 an "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+3.32log (20) =5.32=6(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

2.3.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:

$$Angle of \text{ sector} = \frac{Value of the part}{the whole quantity} * 100$$

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

Men Women Girls Boys

2500 2000 4000 1500

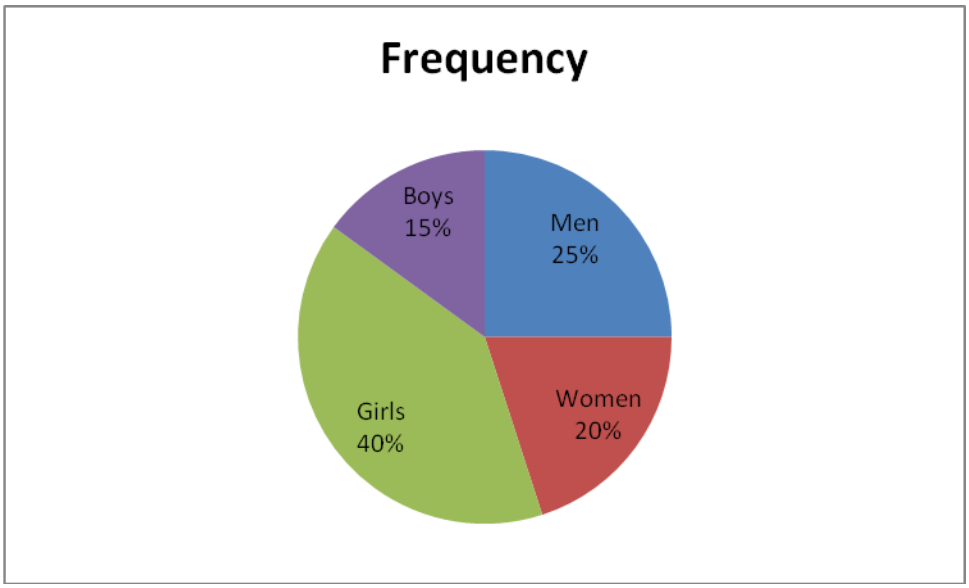
Solutions:

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
- Component or sub divided bar chart.
- Multiple bar charts.

Simple Bar Chart

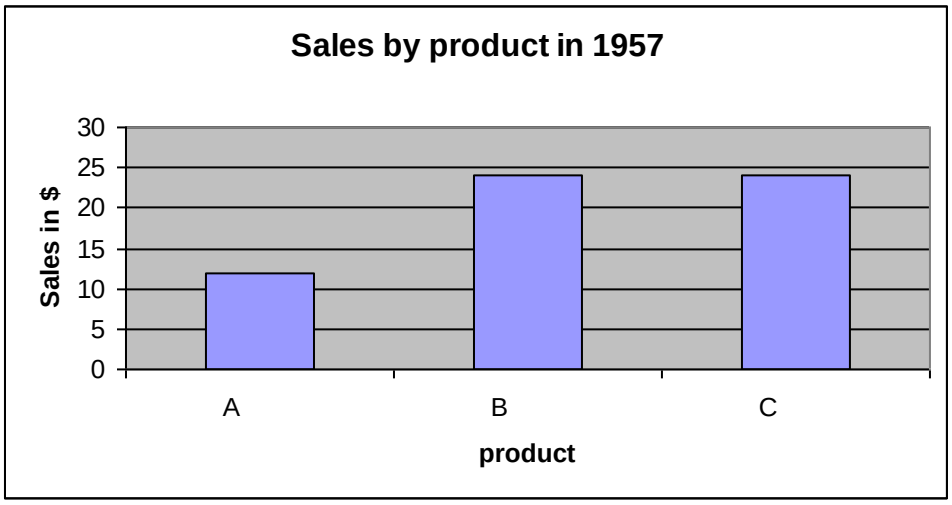
-Are used to display data on one variable.

-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

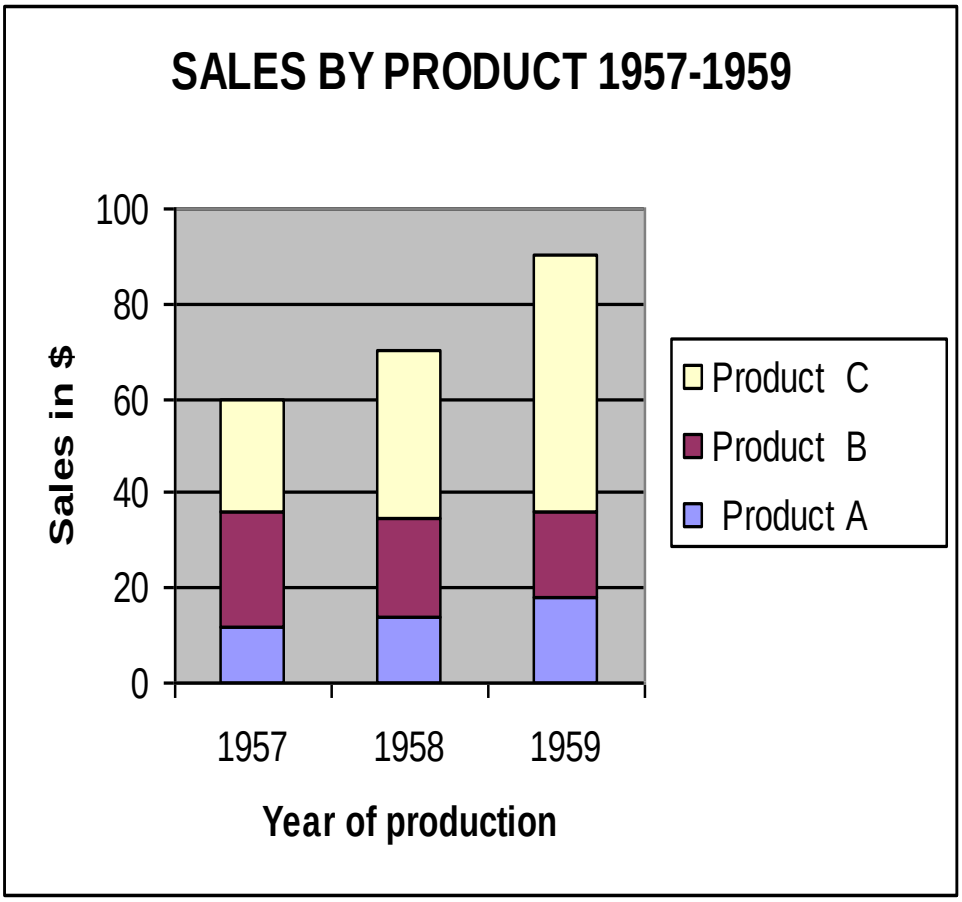
-When there is a desire to show how a total (or aggregate) is divided in to its component parts, we use component bar chart.

-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:



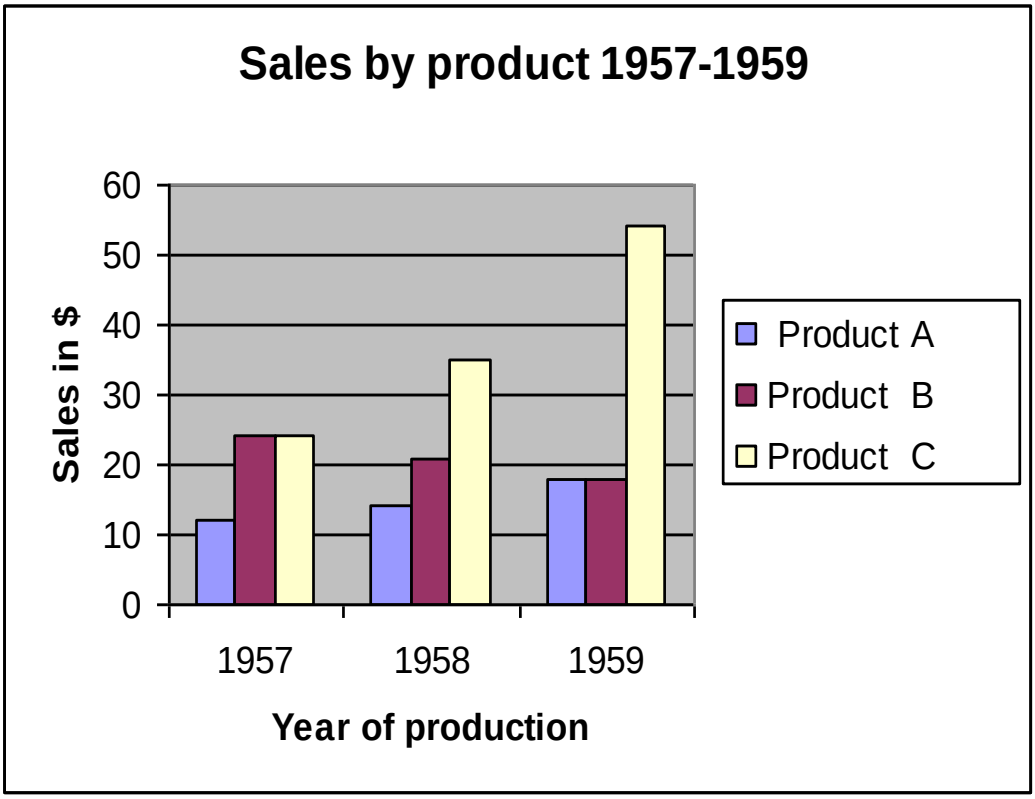
Multiple Bar charts

- 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:



2.4.Graphical Presentation of data

The histogram, frequency polygon and cumulative frequency graph or ogive are most commonly applied graphical representations 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 height 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.

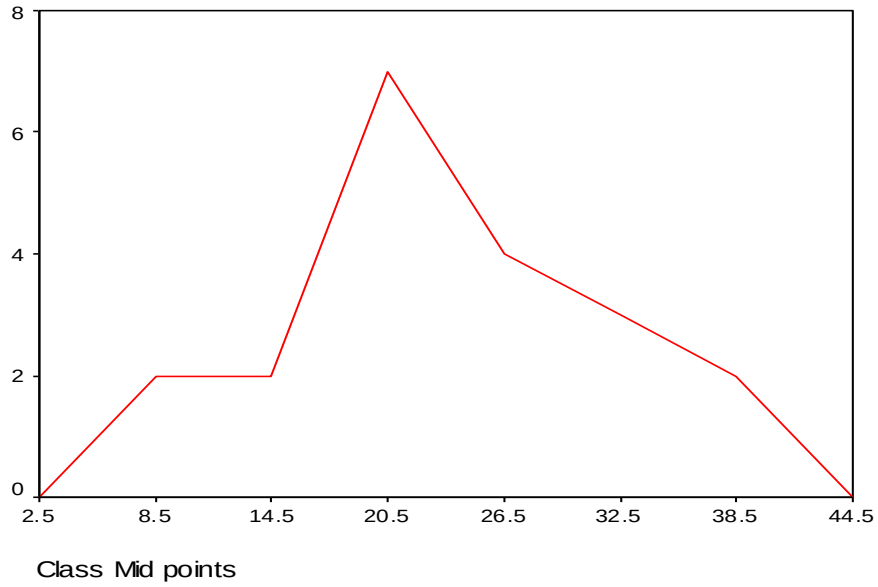
Example: Construct a histogram to represent the previous data (example *).

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 *)

L1. 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. 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.

Objectives:

- ☞ 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 $X_1+X_2+X_3+...+X_N$

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

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. $\sum_{i=1}^{n} (X_i + Y_i) = \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} Y_i$

The sum of the product of the two variables could be written:

$$\sum_{i=1}^{N} (X_i * Y_i) = (X_1 * Y_1) + (X_2 * Y_2) + \ldots + (X_N * Y_N)$$

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

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
- Quantiles (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:

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

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

- 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:

$\mathbf{X_i}$	$\mathbf{f_i}$	$\mathbf{X_i f_i}$
2	2	4
3	1	3
7	3	21
8	1	8
Total	7	36

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

Arithmetic Mean for Grouped Data

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

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

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

Exercises:

1. Marks of 75 students are summarized in the following frequency distribution:

Marks	No. of students
40-44	7
45-49	10
50-54	22
55-59	f_4
60-64	f_5
65-69	6
70-74	3

If 20% of the students have marks between 55 and 59

i. Find the missing frequencies f_4 and f_5.

ii. Find the mean.

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

$$\overline{\mathbf{X}}_c = \frac{\overline{\mathbf{X}}_1\mathbf{n}_1 + \overline{\mathbf{X}}_2\mathbf{n}_2 + \ldots + \overline{\mathbf{X}}_k\mathbf{n}_k}{\mathbf{n}_1 + \mathbf{n}_2 + \ldots\mathbf{n}_k} = \frac{\sum_{i=1}^{k}\overline{\mathbf{X}}_i\mathbf{n}_i}{\sum_{i=1}^{k}\mathbf{n}_i}$$

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 with out repeating the whole process using:

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

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 $X_1, X_2, \ldots, Xn$ 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:

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

$$\mathbf{b).NewMean = -5 * OldMean = -5 * 500 = -2500}$$

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 $X_1, X_2, \ldots X_n$ be the value of items of a series and $W_1, W_2, \ldots W_n$ their corresponding weights , then the weighted mean denoted $\overline{X}_w$ is defined as:

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

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:

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

Merits and Demerits of Arithmetic Mean

Merits:

- It is 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 can not be used in the case of open end classes.
- It can not be determined by the method of inspection.
- It can not be used when dealing with qualitative characteristics, such as intelligence, honesty, beauty.

The Geometric Mean

- The geometric mean of a set of n observation is the n^{th} root of their product.
- The geometric mean of $X_1, X_2, X_3 \ldots X_n$ 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:

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

If observations $X_1, X_2, \ldots X_n$ have weights $W_1, W_2, \ldots W_n$ 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

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

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 modal 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:

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

Where:

$\hat{X} = the\,mode\,of\,the\,distribution$

$w = the\,size\,of\,the\,modal\,class$

$\Delta_1 = f_{mo} - f_1$

$\Delta_2 = f_{mo} - f_2$

$f_{mo} = frequency\,of\,the\,modal\,class$

$f_1 = frequency\,of\,the\,class\,preceeding\,the\,modal\,class$

$f_2 = frequency\,of\,the\,class\ following\,the\,modal\,class$

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:

$45-55$ *is the modal class, since it is a class with the highest frequency.*

$L_{mo} = 45$

$w = 10$

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

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

$f_{mo} = 31$

$f_1 = 29$

$f_2 = 5$

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

$$= 45.71$$

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 $X_1, X_2, \ldots X_n$ 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 $\hat{X}$.

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, 8, 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 is the first cumulative frequency to be greater than or equal to 37.5

$\Rightarrow 50-54$ 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.
- To find Q_i (i=1, 2, 3) we count $\frac{iN}{4}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

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

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,.. D_9 often called the first, the second,…, the ninth deciles respectively.
- To find D_i (i=1, 2,..9) we count $\frac{iN}{10}$ of the classes beginning from the lowest class.
- For grouped data: we have the following formula

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

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 deciles 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 find P_i (i=1, 2,..99) we count $\frac{iN}{100}$ of the classes beginning from the lowest class.

- For grouped data: 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 small 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.

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

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

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

$$N = 493, \quad c = 88, \quad f_{Q_1} = 72$$

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

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

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

ii. Q_2

- determine the class containing the second quartile.

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

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

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

$$N = 493, \quad c = 244, \quad f_{Q_2} = 107$$

$$\Rightarrow Q_2 = L_{Q_2} + \frac{w}{f_{Q_2}}(\frac{2*N}{4} - c)$$
$$= 170 + \frac{10}{72}(246.5 - 244)$$
$$= \underline{\underline{190.23}}$$

iii. Q_3

- determine the class containing the third quartile.

$$\frac{3*N}{4} = 369.75$$
$$\Rightarrow 200 - 210 \quad is\ the\ class containing the third\ quartile.$$

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

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

b) D_7

- determine the class containing the 7^{th} decile.

$$\frac{7*N}{10} = 345.1$$
$$\Rightarrow 190 - 200 \quad is\ the class\ containing the seventh\ decile.$$

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

$$\Rightarrow D_7 = L_{D_7} + \frac{w}{f_{D_7}}(\frac{7*N}{10} - c)$$

$$= 190 + \frac{10}{107}(345.1 - 244)$$

$$= \underline{\underline{199.45}}$$

c) P_{90}

- determine the class containing the 90^{th} percentile.

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

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

$$L_{P_{90}} = 220, \qquad w = 10$$

$$N = 493, \qquad c = 434, \qquad f_{P_{90}} = 3107$$

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

$$= 220 + \frac{10}{31}(443.7 - 434)$$

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

L1.4. Measures of Dispersion (Variation)

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.

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 and coefficient of variation.

4.1.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.

$$R = L - S \qquad , L = l\arg est\ observation$$
$$S = smallest\ observation$$

Range for grouped data:

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

$$R = UCL_k - LCL_1, \qquad UCL_k \ \ is\ upperclass\lim it\ of\ the\ last\ class.$$
$$UCL_1 \ \ is\ lower\ class\lim it\ of\ the\ first\ class.$$

This is some times expressed as:

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

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 can not 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 some times 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

Solution: (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	Freq.
140- 150	17
150- 160	29
160- 170	42
170- 180	72
180- 190	84

190- 200	107	**Solutions:**
200- 210	49	
210- 220	34	In the previous chapter we have obtained the values of
220- 230	31	all quartiles as:
230- 240	16	Q_1= 174.90, Q_2= 190.23, Q_3=203.83
240- 250	12	

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

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

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 ($\overline{X}$):

1. Find the arithmetic mean, $\overline{X}$
2. Find the deviations of each reading from $\overline{X}$.
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:

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

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

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

- 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

$$\overline{X} = 6,\ \tilde{X} = 5.5,\quad \hat{X} = 5$$

Then take the deviations of each observation from these averages.

X_i	4	4	5	5	5	6	7	7	8	9	total
$\lvert X_i - 6 \rvert$	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

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

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

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

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 about the median is always the smallest.

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:

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

$$\Rightarrow C.M.D(\bar{X}) = \frac{M.D(\bar{X})}{\bar{X}} = \frac{1.4}{6} = 0.233 \qquad C.M.D(\tilde{X}) = \frac{M.D(\tilde{X})}{\tilde{X}} = \frac{1.4}{5.5} = 0.255$$

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

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.

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

For the case of frequency distribution it is expressed as:

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

We usually use the following short cut formula.

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

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

4.2.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.

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

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

The following steps are used to calculate the sample variance:

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).

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

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

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

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.

Exercise: 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.

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 alsobe* S

b) $kX_1,\ kX_2, ...kX_n$ *would be* $|k|S$

c) $a + kX_1,\ a + kX_2, ...a + kX_n$ *would be* $|k|\ S$

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 $X_1, X_2, \ldots . X_n$ 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 are 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 = $|k|S = 2*3 = 6$
2. a. They will remain the same.

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

<u>Coefficient of Variation (C.V)</u>

- 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.

Example: 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.

Exercise: 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?

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 deviations 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.

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

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

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

b) Calculate the standard score of A and B

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

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

➔Child B is faster because the time taken by child B is two standard deviations 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.

L1.5. ELEMENTARY PROBABILITY

Definitions of some probability terms

1. **Probability-**is the chance that something will happen.
2. **Experimen**t: probability theory is used as a model for which the outcome occur randomly.
3. **Sample space-** is the set of all possible outcome of the experiment**.** The size of sample space is finite, countable infinite or uncountable infinite .

Examples:

A. the outcome of any of replication is the number of germinating seeds ,since 100 seeds were are monitored the number of germinating could be anything from 0-100.

S= {0,1,2,….100} …… finite sample space

B. The survival time in weeks could be any non _negative integer

S={0,1,2…}…… countable infinite sample space

C. leaf size could be any positive real number $S=R^{+}=\{0,\infty\}$ uncountable infinite sample space

4. **Outcome** :The result of a single trial of a random 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. There are 4 types of events

A. **Null events** – is the empty subset of the sample space.

B. **An atomic event** is a subset consisting of a single element of the sample space.

C. A compound event is a subset consisting more than one element of the sample space.

D. The sample space itself is also an event.

Example: Considering in a rolling of dice only one time, let A be the event of odd numbers, B be the event of even numbers, C be the event of number 8 and D the event of number 4.

$$\Rightarrow A=\{1,3,5\}$$
$$B=\{2,4,6\}$$
$$C=\{\ \}\ or\ empty\ space\ or\ \ impossible\ event$$

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.

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

THE ADDITION RULE :

If E & F are two events which can occur simultaneously then the probability that either E or F will be occur is P(E) +P(F)

The Multiplication Rule: If E & F are two independent events then the probability that both occur P(E) *P(F)

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 $(n_1 * n_2 * \ldots\ldots * n_k)\ ways.$

Example: 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)

1^{st} digit	2^{nd} digit	3^{rd} digit	4^{th} digit
5	5	5	5

There are four steps

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

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

b)

1^{st} digit	2^{nd} digit	3^{rd} digit	4^{th} digit

5	4	3	2

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 4 ways.
3. Selecting the 3rd digit, this can be made in 3 ways.
4. 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 $n!=n*(n-1)*(n-2)*.....*3*2*1$

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

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

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 if two letters are used at a time?

2. How many different permutations can be made from the letters in the word "CORRECTION"?

Solutions: 1. a)

$$Here \quad n=4, \quad there\ are\ four\ disnict\ object$$
$$\Rightarrow There\ are\ 4!=24\ permutations.$$

b)

$$Here \quad n=4, \quad r=2$$
$$\Rightarrow There\ are\ {}_4P_2=\frac{4!}{(4-2)!}=\frac{24}{2}=12\ permutations.$$

2.

$$Here\, n=10$$
$$Of\ \ which\ \ 2\, are\, C,\ \ 2\, are\, O,\ \ 2\, are\ R\ ,1E,1T,1I,1N$$
$$\Rightarrow K_1=2,\ k_2=2,\ k_3=2,\ \ k_4=k_5=k_6=k_7=1$$
$$U\sin g\, the\, 3^{rd}\ \ rule\, of\ \ permutation, there\, are$$
$$\frac{10!}{2!*2!*2!*1!*1!*1!*1!}=453600\ permutations.$$

Combination

A selection of objects with out 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				**Combination**	
AB	BA	CA	DA	AB	BC
AC	BC	CB	DB	AC	BD
AD	BD	CD	DC	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 is 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; and **r**=3

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

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

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

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:

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

Approaches to measuring Probability

There are 3 different conceptual approaches to the study of probability theory. These are:

- The classical approach.
- The frequenters 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

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

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 that
   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.

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

$$\Rightarrow P(A) = \frac{n(A)}{n(S)} = \frac{\binom{30}{0} * \binom{50}{10}}{\binom{80}{10}} = 0.00624$$

- ***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.

Relative frequency, which is the ratio of the occurrence of a singular event and the total number of outcomes**.**

Example: 1. find the relative frequency of the data set

color	frequency	Relative frequency

Purple	7	7/20 = 35%
Blue	3	3/20 = 15%
Pink	5	5/20 = 25%
Orange	5	5/20 = 25%
total	20	20/20 = 100%

2. 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.

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

Subjective approach - this is type of probability based on the beliefs of the person making the probability assessment . subjective probability assessment are often found when events occur only once or at most every few times.

The disadvantages of subjective probability is that two or more person facing the same evidence / problem may arrive different probability. That is for the same problem there may be different decision.

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

   Since the first drawn ball is replaced for a second draw it doesn't affect the second draw. For this reason A and B are independent. Then if we let

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

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

2. Draw a ball with out replacement

This is conditional b/c the first drawn ball is not to be replaced for a second draw in that it does affect the second draw. If we let

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

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

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 by $p(A/B)$ 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

$$given \quad p(A) = 0.25, \qquad p(B) = 0.75, \qquad p(A \cap B) = 0.20$$
$$\text{Re}\,quired \quad p(B/A)$$
$$p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{0.20}{0.25} = 0.80$$

1. 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

$$given \quad p(A) = 0.60, \qquad p(A \cap B) = 0.54$$
$$\text{Re}\,quired \quad p(B/A)$$
$$p(B/A) = \frac{p(A \cap B)}{p(A)} = \frac{0.54}{0.60} = 0.90$$

Exercise: A lot consists of 20 defective 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?

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 $p(A/B) = p(A), \quad P(B/A) = p(B)$

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

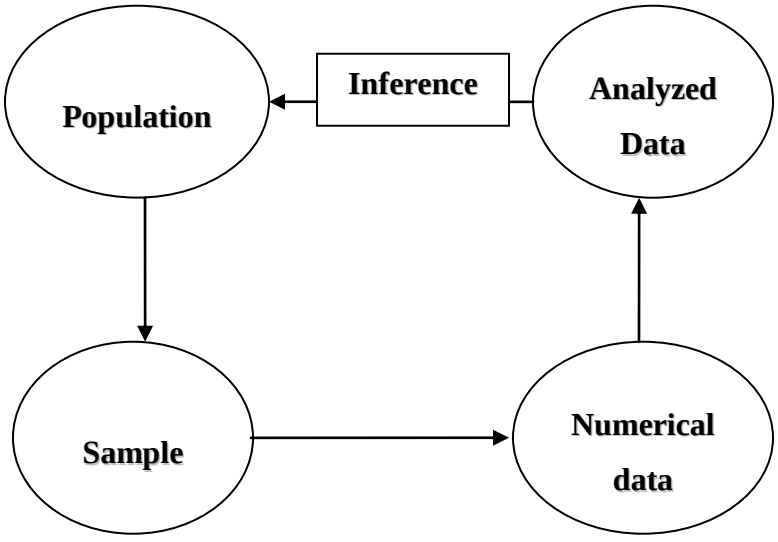
Required $p(A \cap B)$

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

b. $p(A \cap B) = p(A).p(B) = (4/10)(4/10) = 4/25$

L1.6. ESTIMATION AND HYPOTHESIS TESTING

- Inference is the process of making interpretations or conclusions from sample data for the totality of the population.
- It is only the sample data that is ready for inference.
- In statistics there are two ways though which inference can be made.
  - ❖ Statistical estimation
  - ❖ Statistical hypothesis testing.



Data analysis is the process of extracting relevant information from the summarized data.

Statistical Estimation

This is one way of making inference about the population parameter where the investigator does not have any prior notion about values or characteristics of the population parameter.

There are two ways estimation.

1) **Point Estimation**

   It is a procedure that results in a single value as an estimate for a parameter.

2) **Interval estimation**

   It is the procedure that results in the interval of values as an estimate for a parameter, which is interval that contains the likely values of a parameter. It deals with identifying the upper and lower limits of a parameter. The limits by themselves are random variable.

Definitions

Confidence Interval: An interval estimate with a specific level of confidence

Confidence Level: The percent of the time that the true value will lie in the interval estimate given.

Consistent Estimator: An estimator which gets closer to the value of the parameter as the sample size increases.

Degrees of Freedom: The number of data values which are allowed to vary once a statistic has been determined.

Estimator: A sample statistic which is used to estimate a population parameter. It must be unbiased, consistent, and relatively efficient.

Estimate: Is the different possible values which an estimator can assumes.

Interval Estimate: A range of values used to estimate a parameter.

Point Estimate: A single value used to estimate a parameter.

Relatively Efficient Estimator: The estimator for a parameter with the smallest variance.

Unbiased Estimator: An estimator whose expected value is the value of the parameter being estimated.

Point and Interval estimation of the population mean: μ

☞ **Point Estimation**

Another term for statistic is **point estimate**, since we are estimating the parameter value. A **point estimator** is the mathematical way we compute the point estimate. For instance, sum of x_i over n is the point estimator used to compute the estimate of the population means, μ .That is $\overline{X} = \frac{\sum x_i}{n}$ is a point estimator of the population mean.

☞ **Confidence interval estimation of the population mean**

Although $\overline{X}$ possesses nearly all the qualities of a good estimator, because of sampling error, we know that it's not likely that our sample statistic will be equal to the population parameter, but instead will fall into an interval of values. We will have to be satisfied knowing that the statistic is "close to" the parameter. That leads to the obvious question, what is "close"?

We can phrase the latter question differently: How confident can we be that the value of the statistic falls within a certain "distance" of the parameter? Or, what is the probability that the parameter's value is within a certain range of the statistic's value? This range is the confidence interval.

The **confidence level** is the *probability* that the value of the parameter falls within the range specified by the confidence interval surrounding the statistic.

❖ **There are different cases to be considered to construct confidence intervals.**

Case 1: ***If sample size is large or if the population is normal with known variance***

Recall the *Central Limit Theorem,* which applies to the sampling distribution of the mean of a sample. Consider samples of size n drawn from a population, whose mean is μ and standard deviation is σ with replacement and order important. The population can have any frequency distribution. The sampling distribution of $\overline{X}$ will have a mean $\mu_{\bar{x}} = \mu$ and a standard deviation $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, and *approaches a normal distribution as n gets large*. This allows us to use the normal distribution curve for computing confidence intervals.

$$\Rightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad has\,a\,normal\,distribution\,with\,mean = 0\ and\ \text{var}iance = 1$$

$$\Rightarrow \mu = \bar{X} \pm Z\,\sigma/\sqrt{n}$$

$$= \bar{X} \pm \varepsilon, \quad where\,\varepsilon\ is\,a\ measure\,of\ error.$$

$$\Rightarrow \varepsilon = Z\,\sigma/\sqrt{n}$$

- For the interval estimator to be good the error should be small. How it be small?
  - By making n large
  - Small variability
  - Taking Z small

- To obtain the value of Z, we have to attach this to a theory of chance. That is, there is an area of size $1-\alpha$ such that

$$P(-Z_{\alpha/2} < Z < Z_{\alpha/2}) = 1-\alpha$$

$$Where\ \ \alpha = is\,the\ probability\,that\,the\ \ parameter\,lies\,outside\,the\,\text{int}erval$$

$$Z_{\alpha/2} = s\,\text{tan}\,ds\ \ for\,the\,s\,\text{tan}\,dard\ normal\ \text{var}iable\,to\,the\,right\,of\ \ which$$

$$\alpha/2\ probability\,lies, i.e\,P(Z > Z_{\alpha/2}) = \alpha/2$$

$$\Rightarrow P(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1-\alpha$$

$$\Rightarrow P(\bar{X} - Z_{\alpha/2}\,\sigma/\sqrt{n} < \mu < \bar{X} + Z_{\alpha/2}\,\sigma/\sqrt{n}) = 1-\alpha$$

$$\Rightarrow (\bar{X} - Z_{\alpha/2}\,\sigma/\sqrt{n},\ \ \bar{X} + Z_{\alpha/2}\,\sigma/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ \text{int}erval\ \ \ for\,\mu$$

But usually σ^2 is not known, in that case we estimate by its point estimator S^2

$$\Rightarrow (\bar{X} - Z_{\alpha/2}\,S/\sqrt{n},\ \ \bar{X} + Z_{\alpha/2}\,S/\sqrt{n})\ \ is\,a\,100(1-\alpha)\%\ \ conifidence\ \text{int}erval\ \ \ for\,\mu$$

Here are the Z values corresponding to the most commonly used confidence levels.

$100(1-\alpha)$ %	α	$\alpha/2$	$Z_{\alpha/2}$
90	**0.10**	**0.05**	**1.645**

95	**0.05**	**0.025**	**1.96**
99	**0.01**	**0.005**	**2.58**

Case

2: ***If sample size is small and the population variance,*** σ^2 ***is not known.***

$$t = \frac{\bar{X} - \mu}{S/\sqrt{n}} \quad has\ t\ distribution\ with\ n-1\ \deg rees\ of\ freedom.$$

$$\Rightarrow (\bar{X} - t_{\alpha/2}\, S/\sqrt{n},\ \ \bar{X} + t_{\alpha/2}\, S/\sqrt{n})\ \ is\ a\ 100(1-\alpha)\%\ \ confidence\ \text{int}erval\ \ for\ \mu$$

The unit of measurement of the confidence interval is the standard error. This is just the standard deviation of the sampling distribution of the statistic.

Examples:

1. From a normal sample of size 25 a mean of 32 was found .Given that the population standard deviation is 4.2. Find
   a) A 95% confidence interval for the population mean.
   b) A 99% confidence interval for the population mean.

Solution:

a)

$$\bar{X} = 32, \quad \sigma = 4.2, \quad 1-\alpha = 0.95 \Rightarrow \alpha = 0.05, \alpha/2 = 0.025$$

$$\Rightarrow Z_{\alpha/2} = 1.96 \quad from\ \ table.$$

$$\Rightarrow The\ \ required\ \text{int}erval\ \ will\ be\ \bar{X} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$$

$$= 32 \pm 1.96 * 4.2/\sqrt{25}$$

$$= 32 \pm 1.65$$

$$= \underline{\underline{(30.35,\ 33.65)}}$$

b)

$$\bar{X} = 32, \quad \sigma = 4.2, \quad 1-\alpha = 0.99 \Rightarrow \alpha = 0.01,\ \alpha/2 = 0.005$$
$$\Rightarrow Z_{\alpha/2} = 2.58 \quad from \quad table.$$
$$\Rightarrow The \ \ required \text{ int } erval \ \ will \ be \ \bar{X} \pm Z_{\alpha/2}\, \sigma/\sqrt{n}$$
$$= 32 \pm 2.58 * 4.2/\sqrt{25}$$
$$= 32 \pm 2.17$$
$$= \underline{(29.83,\ 34.17)}$$

2. A drug company is testing a new drug which is supposed to reduce blood pressure. From the six people who are used as subjects, it is found that the average drop in blood pressure is 2.28 points, with a standard deviation of .95 points. What is the 95% confidence interval for the mean change in pressure?

Solution:

$$\bar{X} = 2.28, \quad S = 0.95, \quad 1-\alpha = 0.95 \Rightarrow \alpha = 0.05,\ \alpha/2 = 0.025$$
$$\Rightarrow t_{\alpha/2} = 2.571 \ \ with \ df = 5 \quad from table.$$
$$\Rightarrow The \ \ required \text{ int } erval \ \ will \ be \ \bar{X} \pm t_{\alpha/2}\, S/\sqrt{n}$$
$$= 2.28 \pm 2.571 * 0.95/\sqrt{6}$$
$$= 2.28 \pm 1.008$$
$$= \underline{(1.28,\ 3.28)}$$

That is, we can be 95% confident that the mean decrease in blood pressure is between 1.28 and 3.28 points.

Hypothesis Testing

This is also one way of making inference about population parameter, where the investigator has prior notion about the value of the parameter.

Definitions:

- **Statistical hypothesis**: is an assertion or statement about the population whose plausibility is to be evaluated on the basis of the sample data.
- **Test statistic**: is a statistics whose value serves to determine whether to reject or accept the hypothesis to be tested. It is a random variable.

☞ **Statistic test**: is a test or procedure used to evaluate a statistical hypothesis and its value depends on sample data.

There are two types of hypothesis:

Null hypothesis:

- It is the hypothesis to be tested.
- It is the hypothesis of equality or the hypothesis of no difference.
- Usually denoted by H_0.

Alternative hypothesis:

- It is the hypothesis available when the null hypothesis has to be rejected.
- It is the hypothesis of difference.
- Usually denoted by H_1 or H_a.

Types and size of errors:

- Testing hypothesis is based on sample data which may involve sampling and non sampling errors.
- The following table gives a summary of possible results of any hypothesis test:

		Decision	
		Reject H_0	Don't reject H_0
Truth	H_0	Type I Error	Right Decision
	H_1	Right Decision	Type II Error

- **Type I error**: Rejecting the null hypothesis when it is true.
- **Type II error**: Failing to reject the null hypothesis when it is false.

NOTE:

1. There are errors that are prevalent in any two choice decision making problems.
2. There is always a possibility of committing one or the other errors.
3. Type I error (α) and type II error (β) have inverse relationship and therefore, can not be minimized at the same time.

- In practice we set α at some value and design a test that minimize β. This is because a type I error is often considered to be more serious, and therefore more important to avoid, than a type II error.

General steps in hypothesis testing:

1. Specify the null hypothesis (H_0) and the alternative hypothesis (H_1).
2. Specify the significance level, α
3. Identify the sampling distribution (if it is **Z** or **t**) of the estimator.
4. Identify the critical region.
5. Calculate a statistic analogous to the parameter specified by the null hypothesis.
6. Making decision.
7. Summarization of the result.

Hypothesis testing about the population mean, μ :

Suppose the assumed or hypothesized value of μ is denoted by μ_0, then one can formulate two sided (1) and one sided (2 and 3) hypothesis as follows:

1. $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu \neq \mu_0$
2. $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu > \mu_0$
3. $H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu < \mu_0$

Case 1: ***When sampling is from a normal distribution with σ^2 known***

- The relevant test statistic is

$$Z_{cal} = \frac{\overline{X} - \mu_0}{\sigma / \sqrt{n}}$$

- After specifying α we have the following regions (critical and acceptance) on the standard normal distribution corresponding to the above three hypothesis.

Summary table for decision rule:

H_0	Reject H_0 if	Accept H_0 if	Inconclusive if

$\mu \neq \mu_0$	$\lvert Z_{cal} \rvert > Z_{\alpha/2}$	$\lvert Z_{cal} \rvert < Z_{\alpha/2}$	$Z_{cal} = Z_{\alpha/2}\ or\ Z_{cal} = -Z_{\alpha/2}$
$\mu < \mu_0$	$Z_{cal} < -Z_{\alpha}$	$Z_{cal} > -Z_{\alpha}$	$Z_{cal} = -Z_{\alpha}$
$\mu > \mu_0$	$Z_{cal} > Z_{\alpha}$	$Z_{cal} < Z_{\alpha}$	$Z_{cal} = Z_{\alpha}$

Case 2: ***When sampling is from a normal distribution with σ^2 unknown and small sample size***

- The relevant test statistic is

$$t_{cal} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} \sim t \ \ with\ n-1\ \text{degrees of freedom.}$$

- After specifying α we have the following regions on the student t-distribution corresponding to the above three hypothesis.

H_0	Reject H_0 if	Accept H_0 if	Inconclusive if
$\mu \neq \mu_0$	$\lvert t_{cal} \rvert > t_{\alpha/2}$	$\lvert t_{cal} \rvert < t_{\alpha/2}$	$t_{cal} = t_{\alpha/2}\ or\ t_{cal} = -t_{\alpha/2}$
$\mu < \mu_0$	$t_{cal} < -t_{\alpha}$	$t_{cal} > -t_{\alpha}$	$t_{cal} = -t_{\alpha}$
$\mu > \mu_0$	$t_{cal} > t_{\alpha}$	$t_{cal} < t_{\alpha}$	$t_{cal} = t_{\alpha}$

Case 3: ***When sampling is from a non- normally distributed population or a population whose functional form is unknown.***

- If a sample size is large one can perform a test hypothesis about the mean by using:

$$Z_{cal} = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}, \quad if\ \sigma^2\ is\ known.$$

$$= \frac{\overline{X} - \mu_0}{S/\sqrt{n}}, \ if\ \sigma^2\ is\ unknown.$$

- The decision rule is the same as **case I.**

Examples:

1. Test the hypotheses that the average height content of containers of certain lubricant is 10 liters if the contents of a random sample of 10 containers are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3, and 9.8 liters. Use the 0.01 level of significance and assume that the distribution of contents is normal.

Solution:

Let $\mu = Population\ \ mean.\ ,\quad \mu_0 = 10$

Step 1: Identify the appropriate hypothesis

$$H_0 : \mu = 10 \qquad vs \qquad H_1 : \ \mu \neq 10$$

Step 2: select the level of significance, $\alpha = 0.01(given)$

Step 3: Select an appropriate test statistics

t- Statistic is appropriate because population variance is not known and the sample size is also small.

Step 4: identify the critical region.

Here we have two critical regions since we have two tailed hypothesis

$$The\ critical\ \ region\ \ is\ \left|t_{cal}\right| > t_{0.005}(9) = 3.2498$$

$$\Rightarrow (-3.2498,\ 3.2498)\ \ is\ accep\ \tan ce\ \ region.$$

Step 5: Computations:

$$\overline{X} = 10.06,\ \ S = 0.25$$

$$\Rightarrow t_{cal} = \frac{\overline{X} - \mu_0}{S/\sqrt{n}} = \frac{10.06 - 10}{0.25/\sqrt{10}} = 0.76$$

Step 6: **Decision**

Accept H_0 , since $\mathbf{t_{cal}}$ is in the acceptance region.

Step 7: **Conclusion**

At 1% level of significance, we have no evidence to say that the average height content of containers of the given lubricant is different from 10 litters, based on the given sample data.

2. The mean life time of a sample of 16 fluorescent light bulbs produced by a company is computed to be 1570 hours. The population standard deviation is 120 hours. Suppose the hypothesized value

for the population mean is 1600 hours. Can we conclude that the life time of light bulbs is decreasing?

(Use $\alpha = 0.05$ and assume the normality of the population)

Solution:

Let $\mu = Population\ \ mean.\ ,\quad \mu_0 = 1600$

Step 1: Identify the appropriate hypothesis

$$H_0 : \mu = 1600 \qquad vs \qquad H_1 :\ \mu < 1600$$

Step 2: select the level of significance, $\alpha = 0.05\,(given)$

Step 3: Select an appropriate test statistics

Z- Statistic is appropriate because population variance is known.

Step 4: identify the critical region.

$$The critical region\ is\ Z_{cal} < -Z_{0.05} = -1.645$$

$$\Rightarrow (-1.645,\ \infty)\ \ is\ accep\tan ce\ region.$$

Step 5: Computations:

$$Z_{cal} = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} = \frac{1570 - 1600}{120/\sqrt{16}} = -1.0$$

Step 6: **Decision**

Accept H_0, since Z_{cal} is in the acceptance region.

Step 7: **Conclusion**

At 5% level of significance, we have no evidence to say that that the life time of light bulbs is decreasing, based on the given sample data.

Exercise: It is known in a pharmacological experiment that rats fed with a particular diet over a certain period gain an average of 40 gms in weight. A new diet was tried on a sample of 20 rats yielding a weight gain of 43 gms with variance 7 gms. Test the hypothesis that the new diet is an improvement assuming normality.

Test of Association

- Suppose we have a population consisting of observations having two attributes or qualitative characteristics say A and B.
- If the attributes are independent then the probability of possessing both A and B is $P_A * P_B$

Where P_A is the probability that a number has attribute A.

P_B is the probability that a number has attribute B.

- Suppose A has r mutually exclusive and exhaustive classes.

  B has c mutually exclusive and exhaustive classes

- The entire set of data can be represented using $r * c$ contingency table.

B								
A	**B_1**	**B_2**	**.**	**.**	**B_j**	**.**	**B_c**	**Total**
A_1	**O_{11}**	**O_{12}**			**O_{1j}**		**O_{1c}**	**R_1**
A_2	**O_{21}**	**O_{22}**			**O_{2j}**		**O_{2c}**	**R_2**
.								
.								
A_i	**O_{i1}**	**O_{i2}**			**O_{ij}**		**O_{ic}**	**R_i**
.								
.								
A_r	**O_{r1}**	**O_{r2}**			**O_{rj}**		**O_{rc}**	
Total	**C_1**	**C_2**			**C_j**			**n**

- The chi-square procedure test is used to test the hypothesis of independency of two attributes .For instance we may be interested
  - Whether the presence or absence of hypertension is independent of smoking habit or not.
  - Whether the size of the family is independent of the level of education attained by the mothers.
  - Whether there is association between father and son regarding boldness.

- Whether there is association between stability of marriage and period of acquaintance ship prior to marriage.

- The χ^2 statistic is given by:

$$\chi^2_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right] \sim \chi^2_{(r-1)(c-1)}$$

$$Where \;\; O_{ij} = the number of units that belong to category i \; of \; A \; and \; j \; of \; B.$$

$$e_{ij} = Expected \; frequency that belong to category i \; of \; A \; and \; j \; of \; B.$$

- The e_{ij} is given by :

$$e_{ij} = \frac{R_i * C_j}{n}$$

$$Where \;\; R_i = the \; i^{th} row total.$$

$$C_j = the \; j^{th} \; column \; total.$$

$$n = total \; number \; of \; oservations$$

Remark: $n = \sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij} = \sum_{i=1}^{r}\sum_{j=1}^{c} e_{ij}$

- The null and alternative hypothesis may be stated as:

$$H_0 : There is \; no \; association \; between \; A \; and \; B.$$

$$H_1 : not \; H_0 (There is \; association \; between \; A \; and \; B).$$

Decision Rule:Reject H_0 for independency at α level of significance if the calculated value of χ^2 exceeds the tabulated value with degree of freedom equal to $(r-1)(c-1)$.

$$\Rightarrow \text{Reject } H_0 \; if \;\; \chi^2_{cal} = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right] > \chi^2_{(r-1)(c-1)} \;\; at \;\; \alpha$$

Examples:

1. A geneticist took a random sample of 300 men to study whether there is association between father and son regarding boldness. He obtained the following results.

Son		
Father	Bold	Not
Bold	85	59
Not	65	91

Using $\alpha = 5\%$, test whether there is association between father and son regarding boldness.

Solution:

$H_0 : There is\ no\ association\ between\ Father\ and\ Son\ regarding\ boldness$

$H_1 : not\ H_0$

- First calculate the row and column totals

$$R_1 = 144,\quad R_2 = 156,\quad C_1 = 150,\ C_2 = 150$$

- Then calculate the expected frequencies(e_{ij}'s)

$$e_{ij} = \frac{R_i * C_j}{n}$$

$$\Rightarrow e_{11} = \frac{R_1 * C_1}{n} = \frac{144 * 150}{300} = 72 \qquad e_{12} = \frac{R_1 * C_2}{n} = \frac{144 * 150}{300} = 72$$

$$e_{21} = \frac{R_2 * C_1}{n} = \frac{156 * 150}{300} = 78 \qquad e_{22} = \frac{R_2 * C_2}{n} = \frac{156 * 150}{300} = 78$$

- Obtain the calculated value of the chi-square.

$$\chi^2_{cal} = \sum_{i=1}^{2}\sum_{j=1}^{2}\left[\frac{(O_{ij} - e_{ij})^2}{e_{ij}}\right]$$

$$= \frac{(85-72)^2}{72} + \frac{(59-72)^2}{72} + \frac{(65-78)^2}{78} + \frac{(91-78)^2}{78} = 9.028$$

- Obtain the tabulated value of chi-square

$$\alpha = 0.05$$

$$Degrees\ of\ \ freedom = (r-1)(c-1) = 1*1 = 1$$

$$\chi^2_{0.05}(1) = 3.841\ \ from\ table.$$

- The decision is to reject H_0 since $\chi^2{}_{cal} > \chi^2_{0.05}(1)$

Conclusion: At 5% level of significance we have evidence to say there is association between father and son regarding boldness, based on this sample data.

2. Random samples of 200 men, all retired were classified according to education and number of children is as shown below

Education level	***Number of children***		
	0-1	*2-3*	*Over 3*
Elementary	14	37	32
Secondary and above	31	59	27

Test the hypothesis that the size of the family is independent of the level of education attained by fathers. (Use 5% level of significance)

Solution:

H_0 : *There is no association between the size of the family and the level of education attained by fathers.*

H_1 : *not* H_0.

- First calculate the row and column totals

$$R_1 = 83,\ \ R_2 = 117,\ \ \ C_1 = 45,\ C_2 = 96, C_3 = 59$$

- Then calculate the expected frequencies($\mathbf{e_{ij}}$'s)

$$e_{ij} = \frac{R_i * C_j}{n} \quad \Rightarrow e_{11} = 18.675,\ e_{12} = 39.84,\ e_{13} = 24.485$$
$$e_{21} = 26.325,\ e_{22} = 56.16,\ e_{23} = 34.515$$

- Obtain the calculated value of the chi-square.

$$\chi^2_{cal} = \sum_{i=1}^{2}\sum_{j=1}^{3}\left[\frac{(O_{ij}-e_{ij})^2}{e_{ij}}\right]$$

$$= \frac{(14-18.675)^2}{18.675} + \frac{(37-39.84)^2}{39.84} + \ldots + \frac{(27-34.515)^2}{34.515} = 6.3$$

- Obtain the tabulated value of chi-square

$\alpha = 0.05$

$Degrees\ of\ \ freedom = (r-1)(c-1) = 1*2 = 2$

$\chi^2_{0.05}(2) = 5.99\ \ from\ table.$

- The decision is to reject H_0 since $\chi^2_{cal} > \chi^2_{0.05}(2)$

Conclusion: At 5% level of significance we have evidence to say there is association between the size of the family and the level of education attained by fathers, based on this sample data.

L1.7. SIMPLE LINEAR REGRESSION AND CORRELATION

7.1.Introduction

Linear regression and correlation is studying and measuring the linear relation ship among two or more variables. When only two variables are involved, the analysis is referred to as simple correlation and simple linear regression analysis, and when there are more than two variables the term multiple regression and partial correlation is used.

Regression Analysis: is a statistical technique that can be used to develop a mathematical equation showing how variables are related.

Correlation Analysis: deals with the measurement of the closeness of the relation ship which are described in the regression equation.

We say there is correlation if the two series of items vary together directly or inversely.

7.2.Correlation Analysis

Simple Correlation: Suppose we have two variables $X = (X_1, X_2, \ldots X_n)$ and

$$Y = (Y_1, Y_2, \ldots Y_n)$$

- When higher values of X are associated with higher values of Y and lower values of X are associated with lower values of Y, then the correlation is said to be positive or direct.

Examples:

- Income and expenditure
- Number of hours spent in studying and the score obtained
- Height and weight
- Distance covered and fuel consumed by car.

- When higher values of X are associated with lower values of Y and lower values of X are associated with higher values of Y, then the correlation is said to be negative or inverse.

Examples:

- Demand and supply
- Income and the proportion of income spent on food.

The correlation between X and Y may be one of the following

1. Perfect positive (slope=1)
2. Positive (slope between 0 and 1)
3. No correlation (slope=0)
4. Negative (slope between -1 and 0)
5. Perfect negative (slope=-1)

The presence of correlation between two variables may be due to three reasons:

1. One variable being the cause of the other. The cause is called "subject" or "independent" variable, while the effect is called "dependent" variable.
2. Both variables being the result of a common cause. That is, the correlation that exists between two variables is due to their being related to some third force.

Example:

Let X_1= ESLCE result

Y_1= rate of surviving in the University

Y_2= the rate of getting a scholar ship.

Both X_1&Y_1 and X_1&Y_2 have high positive correlation, likewiseY_1 & Y_2 have positive correlation but they are not directly related, but they are related to each other via X_1.

3.Chance: The correlation that arises by chance is called spurious correlation.

Examples:

- Price of teff in Addis Ababa and grade of students in USA.
- Weight of individuals in Ethiopia and income of individuals in Kenya.

Therefore, while interpreting correlation coefficient, it is necessary to see if there is any likelihood of any relation ship existing between variables under study.

The correlation coefficient between X and Y denoted by r is given by

$$r = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2 \sum (Y_i - \bar{Y})^2}} \quad and \quad the shortcut\ formula\ is$$

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{[n\sum X^2 - (\sum X)^2][n\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{[\sum X^2 - n\bar{X}^2][\sum Y^2 - n\bar{Y}^2]}}$$

Remark: Always this r lies between -1 and 1 inclusively and it is also symmetric.

Interpretation of r

1.Perfect positive linear relationship ($if\ r = 1$)

2.Some Positive linear relationship ($if\ r\ \ is$ between 0 and 1)

3.No linear relationship ($if\ r = 0$)

4.Some Negative linear relationship ($if\ r\ \ is$ between -1 and 0)

5.Perfect negative linear relationship ($if\ r = -1$)

Examples:

1. Calculate the simple correlation between mid semester and final exam scores of 10 students (both out of 50)

Student	Mid Sem.Exam (X)	Final Sem.Exam (Y)
1	31	31
2	23	29
3	41	34
4	32	35
5	29	25
6	33	35
7	28	33
8	31	42
9	31	31
10	33	34

Solution:

$n = 10, \quad \bar{X} = 31.2, \quad \bar{Y} = 32.9, \quad \bar{X}^2 = 973.4, \quad \bar{Y}^2 = 1082.4$

$\sum XY = 10331, \quad \sum X^2 = 9920, \quad \sum Y^2 = 11003$

$$\Rightarrow r = \frac{\sum XY - n\bar{X}\bar{Y}}{\sqrt{[\sum X^2 - n\bar{X}^2][\sum Y^2 - n\bar{Y}^2]}}$$

$$= \frac{10331 - 10(31.2)(32.9)}{\sqrt{(9920 - 10(973.4))(11003 - 10(1082.4))}}$$

$$= \frac{66.2}{182.5} = 0.363$$

This means mid semester exam and final exam scores have a slightly positive correlation.

Exercise The following data were collected from a certain household on the monthly income (X) and consumption (Y) for the past 10 months. Compute the simple correlation coefficient.

X:	650	654	720	456	536	853	735	650	536	666
Y:	450	523	235	398	500	632	500	635	450	360

- The above formula and procedure is only applicable on quantitative data, but when we have qualitative data like efficiency, honesty, intelligence, etc we calculate what is called Spearman's rank correlation coefficient as follows:

7.3.Steps

i. Rank the different items in X and Y.
ii. Find the difference of the ranks in a pair , denote them by D_i
iii. Use the following formula

$$r_s = 1 - \frac{6\sum D_i^2}{n(n^2 - 1)}$$

$$Where \quad r_s = coefficient \quad of \ rank \ correlation$$
$$D = the \ difference \ between \ paired \ ranks$$
$$n = the \ number \ of \ pairs$$

Example:

Aster and Almaz were asked to rank 7 different types of lipsticks, see if there is correlation between the tests of the ladies.

Lipstick types	A	B	C	D	E	F	G
Aster	2	1	4	3	5	7	6
Almaz	1	3	2	4	5	6	7

Solution:

X **(R_1)**	**Y** **(R_2)**	**R_1-R_2** **(D)**	**D^2**
2	1	1	1
1	3	-2	4
4	2	2	4
3	4	-1	1
5	5	0	0
7	6	1	1
6	7	-1	1
Total			*12*

$$\Rightarrow r_s = 1 - \frac{6\sum D_i^2}{n(n^2-1)} = 1 - \frac{6(12)}{7(48)} = 0.786$$

Yes, there is positive correlation.

7.4.Simple Linear Regression

- Simple linear regression refers to the linear relation ship between two variables
- We usually denote the dependent variable by Y and the independent variable by X.
- A simple regression line is the line fitted to the points plotted in the scatter diagram, which would describe the average relation ship between the two variables. Therefore, to see the type of relation ship, it is advisable to prepare scatter plot before fitting the model.

- The linear model is:

$$Y = \alpha + \beta X + \varepsilon$$

$$Where: Y = Dependent\,variable$$

$$X = independent\,variable$$

$$\alpha = \text{Re}\,gression\,constant$$

$$\beta = regression\,slope$$

$$\varepsilon = random\,disturbance\,term$$

$$Y \sim N(\alpha + \beta X, \sigma^2)$$

$$\varepsilon \sim N(0, \sigma^2)$$

- To estimate the parameters ($\alpha\ and\ \beta$) we have several methods:
  - The free hand method
  - The semi-average method
  - The least square method
  - The maximum likelihood method
  - The method of moments
  - Bayesian estimation technique.

- The above model is estimated by: $\hat{Y} = a + bX$

Where a is a constant which gives the value of Y when X=0 .It is called the Y-intercept. b is a constant indicating the slope of the regression line, and it gives a measure of the change in Y for a unit change in X. It is also regression coefficient of Y on X.

- a and b are found by minimizing $SSE = \sum \varepsilon^2 = \sum (Yi - \hat{Y}_i)^2$

$$Where: Y_i = observed\,value$$

$$\hat{Y}_i = estimated\,value = a + bX_i$$

And this method is known as OLS (ordinary least square)

- Minimizing $SSE = \sum \varepsilon^2$ gives

$$b = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum X^2 - n\bar{X}^2}$$

$$a = \bar{Y} - b\bar{X}$$

Example 1: The following data shows the score of 12 students for Accounting and Statistics examinations.

a) Calculate a simple correlation coefficient
b) Fit a regression equation of Statistics on Accounting using least square estimates.
c) Predict the score of Statistics if the score of accounting is 85.

	Accounting X	**Statistics Y**	**X^2**	**Y^2**	**XY**
1	74.00	81.00	5476.00	6561.00	5994.00
2	93.00	86.00	8649.00	7396.00	7998.00
3	55.00	67.00	3025.00	4489.00	3685.00
4	41.00	35.00	1681.00	1225.00	1435.00
5	23.00	30.00	529.00	900.00	690.00
6	92.00	100.00	8464.00	10000.00	9200.00
7	64.00	55.00	4096.00	3025.00	3520.00
8	40.00	52.00	1600.00	2704.00	2080.00
9	71.00	76.00	5041.00	5776.00	5396.00

10	33.00	24.00	1089.00	576.00	792.00
11	30.00	48.00	900.00	2304.00	1440.00
12	71.00	87.00	5041.00	7569.00	6177.00
Total	687.00	741.00	45591.00	52525.00	48407.00
Mean	57.25	61.75			

a)

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{n\sum X^2 - \left(\sum X\right)^2} \times \sqrt{n\sum Y^2 - \left(\sum Y\right)^2}}$$

$$r = \frac{12 \times 48407 - 687 \times 741}{\sqrt{12 \times 45591 - 687^2} \times \sqrt{12 \times 52525 - 741^2}}$$

$$r = \mathbf{0.9194}$$

The Coefficient of Correlation (r) has a value of 0.92. This indicates that the two variables are positively correlated (Y increases as X increases).

b)

$$b = \frac{\sum XY - n\overline{X}\,\overline{Y}}{\sum X^2 - n\overline{X}^2}$$

$$b = \frac{48407 - 12 \times 57.25 \times 61.75}{45591 - 12 \times (57.25)^2}$$

$$b = 0.9560$$

$$a = \overline{Y} - b\overline{X}$$

$$a = 61.75 - 0.9560 \times 57.25$$

where: $a = 7.0194$

$$\Rightarrow \quad \hat{Y} = 7.0194 + 0.9560X \quad is\,the\,estimated\,regression\,line.$$

c) Insert X=85 in the estimated regression line.

$$\hat{Y} = 7.0194 + 0.9560X$$

$$= 7.0194 + 0.9560(85) = 88.28$$

Exercise: A car rental agency is interested in studying the relationship between the distance driven in kilometer (Y) and the maintenance cost for their cars (X in birr). The following summarized information is given based on samples of size 5.

$$\sum_{i=1}^{5} X_i^{\,2} = 147{,}000{,}000 \qquad \sum_{i=1}^{5} Y_i^{\,2} = 314$$

$$\sum_{i=1}^{5} X_i = 23{,}000\,, \qquad \sum_{i=1}^{5} Y_i = 36, \qquad \sum_{i=1}^{5} X_i\, Y_i = 212,\,000$$

a) Find the least squares regression equation of Y on X
b) Compute the correlation coefficient and interpret it.
c) Estimate the maintenance cost of a car which has been driven for 6 km

- To know how far the regression equation has been able to explain the variation in Y we use a measure called coefficient of determination (r^2)

$$i.e \quad r^2 = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

$$Where \;\; r = the\, simple\, correlation\; coefficient.$$

- r^2 gives the proportion of the variation in Y explained by the regression of Y on X.
- $1 - r^2$ gives the unexplained proportion and is called coefficient of indetermination.

Example: For the above problem (example 1): $r = 0.9194$

$\Rightarrow r^2 = 0.8453 \Rightarrow$ 84.53% of the variation in Y is explained and only 15.47% remains unexplained and it will be accounted by the random term.

- Covariance of X and Y measures the co-variability of X and Y together. It is denoted by S_{XY} and given by

$$S_{XY} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum XY - n\bar{X}\bar{Y}}{n-1}$$

- Next we will see the relation ship between the coefficients.

i. $r = \frac{S_{XY}}{S_X S_Y} \Rightarrow r^2 = \frac{S_{XY}^2}{S_X^2 S_Y^2}$

ii. $r = \frac{bS_X}{S_Y} \Rightarrow b = \frac{rS_Y}{S_X}$

- When we fit the regression of X on Y , we interchange X and Y in all formulas, i.e. we fit

$$\hat{X} = a_1 + b_1 Y$$

$$b_1 = \frac{\sum XY - n\bar{X}\bar{Y}}{\sum Y^2 - n\bar{Y}^2}$$

$$a_1 = \bar{X} - b_1\bar{Y} \quad , \quad r = \frac{b_1 S_Y}{S_X}$$

Here X is dependent and Y is independent.

7.5.Choice of Dependent and Independent variable

- In correlation analysis there is no need of identifying the dependent and independent variable, because r is symmetric. But in regression analysis

If b_{YX} is the regression coefficient of Y on X

b_{XY} is the regression coefficient of X on Y

Then $r = \frac{b_{YX} S_X}{S_Y} = \frac{b_{XY} S_Y}{S_X} \Rightarrow r^2 = b_{YX} * b_{XY}$

- Moreover, b_{YX} and b_{XY} are completely different numerically as well as conceptually.

- Let us consider three cases concerning these coefficients.

1. If the correlation is perfect positive, i.e. $r = 1$ then the b values reciprocals of each other.

2. *If* $S_X = S_Y$, then irrespective of the value of r the b values are equal, i.e. $r = b_{YX} = b_{XY}$ (but this is unlikely case)
3. The most important case is when $S_X \neq S_Y$ *and* $r \neq 1$, here the b values are not equal or reciprocals to each other, but rather the two lines differ , intersecting at the common point ($\overline{X}, \overline{Y}$)
    - Thus to determine if a regression equation is X on Y or Y on X , we have to use the formula $r^2 = b_{YX} * b_{XY}$
    - If $r \in [-1,1], then \ \ our\ assumption\ is\ \ correct$
    - If $r \notin [-1,1], then \ \ our\ assumption\ is\ \ wrong$

Example: The regression line between height (X) in inches and weight (Y) in lbs of male students are:

$$4Y - 15X + 530 = 0 \ and$$
$$20X - 3Y - 975 = 0$$

Determine which is regression of Y on X and X on Y

Solution

We will assume one of the equation as regression of X on Y and the other as Y on X and calculate r

$$Assume \ \ 4Y - 15X + 530 = 0 \ is\, regression of \ X \ on \ Y$$
$$20X - 3Y - 975 = 0 \ \ is\, regression of \, Y \, on \ X$$

Then write these in the standard form.

$$4Y - 15X + 530 = 0 \Rightarrow X = \frac{530}{15} + \frac{4}{15}Y \Rightarrow b_{XY} = \frac{4}{15}$$
$$20X - 3Y - 975 = 0 \Rightarrow Y = \frac{-975}{3} + \frac{20}{3}X \Rightarrow b_{YX} = \frac{20}{3}$$

$$\Rightarrow r^2 = b_{XY} * b_{YX} = \left(\frac{4}{15}\right)\left(\frac{20}{3}\right) = 1.78 > 1,$$

This is impossible (contradiction). Hence our assumption is not correct. Thus

$$4Y - 15X + 530 = 0 \; is\, regression of\; Y\; on\; X$$

$$20X - 3Y - 975 = 0 \; is\, regression of\; X\; on\; Y$$

To verify:

$$4Y - 15X + 530 = 0 \Rightarrow Y = \frac{-530}{4} + \frac{15}{4}X \Rightarrow b_{YX} = \frac{15}{4}$$

$$20X - 3Y - 975 = 0 \Rightarrow X = \frac{975}{20} + \frac{3}{20}Y \Rightarrow b_{XY} = \frac{3}{20}$$

$$\Rightarrow r^2 = b_{YX} * b_{XY} = \left(\frac{15}{4}\right)\left(\frac{3}{20}\right) = \frac{9}{16} \in [0,1]$$