Basic Statistics

Meaning of Statistics:

The word Statistics has basically three meanings:

1. Statistics is the quantitative information of any inquiry. For example, the statistics of birth and death, the statistics of export and import, the statistics of the evolution of human races, the statistics of the products of all human activity in society, the statistics of manpower, loss and profit of different organizations.
2. It is scientific techniques of collection, analysis and interpretation of quantitative data. For example, the method of collection of data related to birth and family planning adoption and then interpretation regarding population growth.
3. It is also used in singular sense to explain the population unknown characteristics by a function of sample observations. This function is known as **Statistic**.

Definition of Statistics:

Statistics is concerned with scientific methods for **collecting, organizing, summarizing, presenting and analyzing** sample data as well as drawing valid conclusions about population characteristics and making reasonable decisions on the basis of such analysis.

According to *Lovitt*, Statistics is the science which deals with collection, classification and tabulation of numerical facts as the basis for explanation, description and comparison of phenomenon.

Characteristics/ Salient features of Statistics:

Statistics should possess the following characteristics:

(i) Statistics should deal with aggregate of individuals rather than with individual alone.
(ii) Statistics should be expressed as numerical figures.
(iii) Statistics should have the property of being varied by multiplicity of causes.
(iv) Statistics should be collected with reasonable standards of accuracy.
(v) Statistics should be obtained for pre-determined purposes.
(vi) Statistics data should be collected with a view to make comparison with other data.

Scope and Use of Statistics:

The scope and use of statistics are so wide and universal that they cannot be enumerated instantly in a few words. Statistics is used in researches of almost all disciplines. In fact, there is hardly any field where statistical methods cannot be applied:

1) **Statistics and Physics:** Applying a statistical approach to thermodynamics can lead to a deeper understanding concepts such as temperature and entropy.
2) **Statistics and economics:** Statistical data and techniques of statistical analysis have to immensely useful involving economical problem. Such as wages, price, time series analysis, demand analysis.

Dr. Md. Atiqul Islam, Department of Statistics, JnU, Dhaka
E-mail: atique@stat.jnu.ac.bd

3) **Statistics and business:** Statistics-is an irresponsible tool of production control. Busine executive are relying more and more on statistical techniques for studying the much and desire of the valued customers.
4) **Statistics and industry:** In industry statistics is widely used inequality control. In production engineering to find out whether the product is confirming to the specifications or not. Statistical tools, such as inspection plan, control chart etc.
5) **Statistics and mathematics:** Statistics are intimately related recent advancements in statistical technique are the outcome of wide applications of mathematics.
6) **Statistics and modern science:** In medical science, the statistical tools for collection, presentation and analysis of observed facts relating to causes and incidence of dieses and the result of application various drugs and medicine are of great importance.
7) **Statistics and war:** In war, the theory of decision function can be a great assistance to the military and personal to plan "maximum destruction with minimum effort."

Limitation of Statistics:

The drawbacks of the statistics are:

(i) Statistics is not suited to the study of qualitative phenomenon.
(ii) Statistics does not study individuals.
(iii) Statistical laws are not exact.
(iv) Statistics is liable to misused.

Types of Statistics:

Statistics deals with both statistical data and statistical methods. Statistical methods are again divided into two branches like

(i) Descriptive Statistics and
(ii) Inferential Statistics

Descriptive Statistics: CTPA

Descriptive Statistics deals with collection, tabulation, presentation and analysis of data without considering the theory of probability. The study of frequency distribution is an aspect of tabulation. The analytical aspects deal with the measures of central tendency, measures of dispersion, skewness and kurtosis. The shape of the frequency curve is studied by skewness and kurtosis. All the above measures are used for univariate, bi-variate and multivariate data. The study of correlation, regression and association of attributes are included in the bi-variate descriptive statistics.

Inferential Statistics:

Statistics is based on inductive logic. Inferential Statistics is concerned with making estimates, predictions and generalizations, or reaching decisions about population based on sample observations. The method of taking decision is known as statistical inference. The inference is made by sampling, sampling distribution, estimation of parameter and test regarding any hypothesis on parameter.

Dr. Md. Atiqul Islam, Department of Statistics, JnU, Dhaka
E-mail: atique@stat.jnu.ac.bd

Statistical data:

Any measurement of one or more characteristics recorded (as a result of observation, interview and so on) either from population or sample units are called data. Data are the raw, disorganized facts and figures collected from any field of inquiry. Data can be **numerical** (e.g. temperature) or **non-numerical** (e.g., having cancer).

For example, the heights of 14 randomly selected persons from a group of N = 100 persons are as follows: 152, 160, 158, 155, 150, 152, 151, 150, 153, 154, 153, 154, 151, 155. This information on height of people constitutes the data.

Types of data

Statistical data depending upon the sources are of two types, they are:

(i) Primary data
(ii) Secondary Data

Primary Data:

The data, which are collected from the main sources by basic investigation or direct observation of the experimental units, are called primary data.

Secondary data:

The data that are collected from indirect sources such as from any institution or organization, publication, report, journal etc. is called secondary data.

Collected data are stored in two fashions, they are:

(i) Raw data
(ii) Classified data

Raw data:

The data which are collected from sampling units and stored or recorded without any systematic fashion are known as raw data.

For example: The number of road accident in 5 selected days in highways in a year: 12, 10, 4, 12, 5.

Classified data:

The primary data which are presented in a systematic fashion in rows or columns or even in ordered way are known as classified data.

For example, the following data represent the number of some selected private universities according to their number of students:

C.I. of number of Students	Number of Universities
<500	5
500-1000	15
1000-1500	4
1500-2000	3
2000+	2

Continuous Data:

A set of data is said to be continuous if the observations may take on any value within a finite or infinite interval. For example, height, weight, temperature, the amount of sugar in an orange, the time required to run a mile.

Categorical Data:

A set of data is said to be categorical if the values or observations can be sorted according to category. For example, people have the characteristic of 'gender' with categories 'male' and 'female'.

Sources of Statistical data:

Statistical data may be collected in a variety of ways. These sources may broadly be categorized as primary source and secondary source. Primary data come mainly from direct field operations, which may either be a census or a specially designed survey. On the other hand, secondary data are usually procured from already published or unpublished documents rather than undertaking first-hand field investigations. So, the primary data collected by an agency or organization, constitute the secondary data in the hands of other agencies. Bangladesh Bureau of Statistics (BBS), for example, conducts occasional surveys on various aspects, such as health, migration, marriage and morbidity. Such data in their hands are regarded as primary data. They are compiling, analyzing and preparing periodic reports on the issues. If these data are used by some other interested groups to serve their own purpose, the BBs data become secondary in nature to them.

Scale of Measurement/ Measurement of Data:

Nominal Scale:

Categorical data and numbers that are simply used as identifiers or names represent a nominal scale of measurement. A set of data is said to be nominal if the observations can be assigned a code in the form of a number where the numbers are simply labels. One can count but not order or measure nominal data. For example, in a dataset, male could be coded as 0, female as 1; marital status of an individual could be coded as Y if married, N if single; Also, numbers on the back of a football jersey (Leo 10 = Leonel Messi) are examples of nominal data.

Ordinal Scale:

An ordinal scale of measurement represents an ordered series of relationships or rank order.
A set of data is said to be ordinal if the observations can be ranked (put in order) or have a rating scale attached. One can count and order, but not measure ordinal data. For example, suppose a group of people were asked to taste varieties of biscuit and classify each biscuit on a rating scale of 1 to 5, representing strongly dislike, dislike, neutral, like, strongly like. A rating of 5 indicates more enjoyment than a rating of 4, for example, so such Likert-type scale data are ordinal.

Interval Scale:

An interval scale is a scale of measurement where the distance between any two adjacent units of measurement (or 'intervals') is the same but the zero point is arbitrary. Scores on an interval scale can be added and subtracted but cannot be meaningfully multiplied or divided. For example, the time interval between the start of years 1981 and 1982 is the same as that between 1983 and 1984, namely

365 days. The zero-point, year 1 AD, is arbitrary; time did not begin then. The Celsius scale is a clear example of the interval scale of measurement. Thus, 0 degree Celsius is interval data.

Ratio Scale:

The ratio scale of measurement is similar to the interval scale in that it also represents quantity and has equality of units. Ratio data is one, which can take numeric values that are actual as well as absolute. The zero value on this scale is absolutely zero. For example, physical measures will represent ratio data (for example, height and weight). If one is measuring the length of a piece of wood in centimeters, there is quantity, equal units, and that measure cannot go below zero centimeters. A negative length is not possible.

Comparison of measurement Scale

To give a better overview the values in 'Mathematical Operators', 'Advanced operations' and 'Central tendency' are only the ones this level of measurement introduces.

Incremental progress	Measure property	Mathematical operators	Advanced operations	Central tendency
Nominal	Classification, membership	$=, \neq$	Grouping	Mode
Ordinal	Comparison, level	$>, <$	Sorting	Median
Interval	Difference, affinity	$+, -$	Yardstick	Mean, Deviation
Ratio	Magnitude, amount	$\times, /$	Ratio	Geometric mean, coefficient of variation

Population

An aggregate of all individuals or items under investigation defined on some common characteristics is called a population.

For example, 1st year 2nd Semester B.Sc. (honours) students in Physics (session: 2020-2021) of JnU constitute a population. Here, the *common characteristics* are:

(i) Students of JnU

(ii) Students of 1st year 2nd Semester B.Sc. (honours) students in Physics and

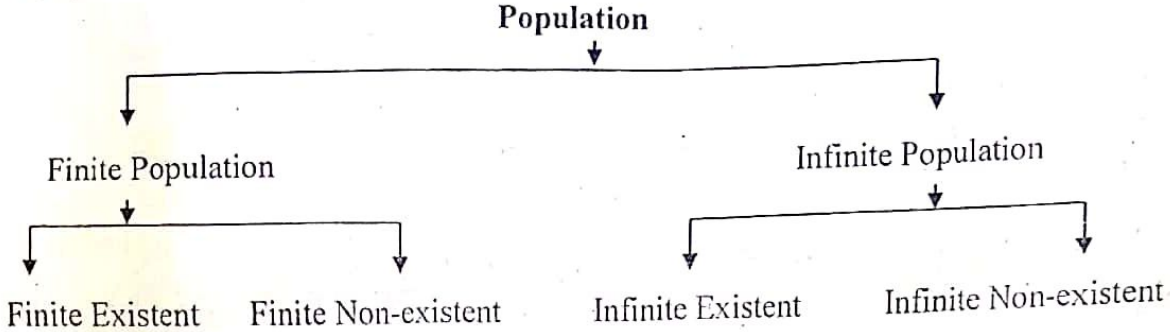
(iii) Students of the session 2020-2021

Or, an aggregate of all individuals or items under investigation according to some pre-determined objective and are available in a specified area at a specified time period.

For example, if the objective is to estimate the per capita salary in 2018 of female employees working in different garments industries in Bangladesh, all female employees in all industries of Bangladesh during a particular time period constitute the population.

Types of Population:

A population can be classified in different types which is shown by the following diagram:



Finite Population:

A population consisting of a finite number of individuals or items is called a finite population. For example, 1st year 2nd Semester B.Sc. (honours) students in Physics (session: 2020-2021) of JnU constitute a finite population.

Infinite Population:

A population consisting of an infinite number of individuals or items is called an infinite population. For example, if we toss a coin for an infinite number of times and write down the upturned face of the coin then the sequence of Head (H) and Tail (T) (e.g., HHHTTHT-----) will constitute an infinite population.

Sample

A small but representative part with finite number individuals or items of a population which is under investigation is called a sample.

For example, a group of students, representing the 1st year 2nd semester B.Sc. (honours) students of JnU is called a sample.

Random Sample:

If each individual or item in the population from which a sample has been drawn or selected, has an equal chance of being included in the sample, then the sample is called a random sample.

For example, If we have a complete list of 100 students and if we select a sample of 20 students from these 100 students completely at random, then each of the students has an equal chance of being included in the sample. Therefore, the sample 20 students is a random sample.

Variable:

A variable is a characteristic whose value can vary from person to person, object to object or from phenomenon to phenomenon.

For example, (i) gender is variable which is composed of two categories, male and female and it varies from one to another, (ii) age is a variable which may vary from person to person and may assume values 10 years, 15 years 20 years, (iii) level of arsenic in water may vary between wells and so on.

Types of variable:

There are two types of variables, they are

(i) Qualitative variable: A qualitative variable is one for which numerical measurement is not possible, such as hair colour, religion, race, sex, stages of cancer, smoking status etc.

(ii) Quantitative variable: A quantitative variable is one for which the resulting observations are numeric and thus possesses a natural ordering. Example: Age, Height, Family size etc.

Quantitative variable can also be classified as

(i) Discrete variable: When the variable can assume only the isolated values, the variable is called discrete variable. Example: the number of children in a family, fishes in a pond etc.

(ii) Continuous variable: A variable is said to be continuous if it assumes any value within a certain range. Example: age, height, weight, blood count, temperature, rainfall, level of arsenic in water, count of bacteria etc.

Attribute: A qualitative characteristics, when used to classify a series or individuals into two or more mutually exclusive and exhaustive classes is called an attribute. Example: sex (male/female), hair colour (black/gray), smoking (smoker/non-smoker/never smoker) etc.

Condensation of Data: The primary data which are collected through survey are called raw data. The raw data are not always suitable for proper statistical analysis. For analytical purpose and for the purpose of comprehensive idea about the population under investigation the data are usually ordered, classified and tabulated. The process of ordering, classification and tabulation is called condensation of data.

Array: The quantitative values of a variable are arranged in ascending or in descending order of magnitude. This arrangement of values of a variable is called an array.

Classification: Classification is a technique by which data are divided into several related classes consecutively according to some characteristics so that data of each class are homogeneous in nature. For example, different administrative divisions of Bangladesh can be classified according to area (in square kilometer) as follows

Table: Area in square in different administrative divisions of Bangladesh

Division	Area (in square kilometer)
Barisal	13,225.20
Chittagong	33,908.55
Dhaka	20,593.74
Khulna	22,284.22
Mymensingh	10,584.06
Rajshahi	18,153.08
Rangpur	16,184.99
Sylhet	12,635.22

Tabulation: Tabulation is the process of arranging the data in an orderly manner into rows and columns. A statistical table is the logical listing of related quantitative information in vertical columns and in horizontal rows with sufficient explanatory and qualifying words, phrases and statements in the form of title, sub-titles, headings and notes to make clear the full meaning of data and their origin.

The primary difference between classification and tabulation is that the process of classifying data into groups is known as classification of data, whereas tabulation is the act of presenting data in tabular form, for better interpretation.

References

1. Hoel, P. G., *Introduction to Mathematical Statistics*, 5th Ed, John Wiley, NY
2. Islam M. N., *An Introduction to Statistics and Probability*, 3rd Edition
3. Mostafa M. G., *Methods of Statistics.*
4. Bhuyan, K. C., *Methods of Statistics.*
5. Bulmer, M. G., *Principles of Statistics*, 2nd Ed, Oliver and Boyd, Edinburgh
6. Shil, R.N. and Debnath, S.C. *An Introduction to the Theory of Statistics*
7. Rahman, M.S. *Statistics and Probability: An Introductory Analysis*

SOME DEFINITIONS REGARDING FREQUENCY DISTRIBUTION

FREQUENCY DISTRIBUTION:

A frequency distribution is a table in which the values for a variable are grouped into classes and observed frequencies are recorded.

CLASS:

In the process of condensation, raw data are assigned to some chosen groups of appropriate size. These groups are called *classes*. A class is thus an interval containing observations, each observation being classified into one and only one class.

FREQUENCY:

The number of observations or values falling into each group or class is called *class frequency* or simply *frequency*. The frequency thus shows how many times a particular value or observation is repeated. For example, if a value of 10 occurs 6 times in a data set, then 6 is the frequency of 10.

CLASS INTERVAL:

For numerical data, the frequencies of a particular class are bounded by two values. The width or length of the class formed by these two boundary values is known as the class interval.

CLASS LIMITS:

The smallest value of a class is technically known as the ***lower-class limit*** of the interval, while the largest value is known as the ***upper-class limit*** of the interval. Thus, for a class interval 15-19, 15 is the lower limit and 19 is the upper limit.

CLASS MID-POINT:

The mid-point or mid-value of a class is the value that falls in the middle of the class interval and is obtained as the average of the two class limits. For the class interval 15-19, the mid-point is 17.

CLASS WIDTH:

The size of a class is referred to as *class width* and is the difference between the two class limits. For example, a class with interval 45-50 has a class width of 5.

OPEN INTERVAL:

An open interval is an interval with one of its limits (on either side) indeterminate. Thus, an age of a person recorded as less than 45 years (i.e., < 45) also forms an open interval.

CLASS BOUNDARY:

In an inclusive method, it is necessary to make an adjustment to determine the correct class intervals and to have continuity. The adjustment consists in finding a Correction Factor (CF). The Correction Factor can be expressed as

$$CF = \frac{1}{2}(\text{Lower limit of the second class} - \text{Upper limit of the first class})$$

For each class, this CF is subtracted from the lower limit and added to the upper limit to maintain the continuity of data. Thus, obtaining a true class interval is called a class boundary. For example, if the class limits are 20 and 29, then all values between 19.5 and 29.5 would actually fall in the given class, so the class boundaries are 19.5 and 29.5.

TALLY MARKS:

To indicate the accommodation of an observation to a particular class, a tally sign (/) is used. This sign is known as a tally mark. The tally mark facilitates the count of frequencies of a class.

CUMULATIVE FREQUENCY:

Cumulative frequency is computed by adding successive class frequencies from top to bottom. (The entry corresponding to the top interval is the frequency of that class; the entry opposite the second interval is the sum of the frequencies in the first- and second-class intervals etc.)

INCLUSIVE METHOD:

Under the inclusive method, the upper limit of one class is included in that class itself. The method is inclusive in the sense that it includes both ends of the intervals so that its inclusion does not alter the width of the interval.

EXCLUSIVE METHOD:

When the class intervals are so fixed, the upper limit of one class is the lower limit of the next class. This type of classification is conventionally known as the exclusive method.

STEPS IN CONSTRUCTION OF FREQUENCY DISTRIBUTION:

Following are the steps for the construction of a frequency distribution:

1) Find out ***Range (R)*** by subtracting the lowest value from the highest value of a variable.
2) Take decision regarding the number of classes and class interval:

   Let k = the number of classes in a frequency table. The number of classes k should not be less than 4 and should not be more than 20. However, the value of k can be found by a mathematical formula where

$$k = 1 + 3.322 log_{10} N,$$

   here, N is the total number of observations. (This rule for k is known as **Sturge's** Rule for the number of classes).

   Once the value of k is decided, ***the interval (h, width) of a class*** is found out by

$$h = \frac{Range}{k}$$

3) Arrange the table with three columns having headings: ***Class Interval, Tally Marks and Frequency***. The first class interval will start with the smallest value and continue until the interval with the highest value of the given series of data is reached.
4) Read the items and give a tick mark or circle to each of the values of the original table of raw data and put a tally mark against the appropriate class interval. It is convenient to mark each fifth by a diagonal (\). Thus, exhaust all the values one after another.
5) Count the number of tally marks corresponding to each class interval and write the result in the respective frequency column.

EXERCISE:

Numbers obtained in Statistics course by 40 students of 1st year 2nd semester is given below:

40	38	44	28	30	22	35	42	40	36
50	67	25	58	53	48	65	35	55	39
72	44	70	55	62	20	78	46	57	68
59	34	41	56	60	42	64	73	38	41

Construct a suitable frequency distribution

SOLUTION:

We have a total number of observations, N = 40

Lowest observation = 20

Highest observation = 78

Therefore, $Range = Highest\ observation - Lowest\ observation = 78 - 20 = 58$

The number of classes, $k = 1 + 3.322 log_{10} N$ [10 base log]

$= 1 + 3.322 log_{10} 40$

$= 6.32 \approx 6$

Hence, the class Interval,

$$h = \frac{Range}{k} = \frac{58}{6} = 9.67 \approx 10$$

Thus, we can construct the following frequency table:

Table 1: Grouped Frequency Distribution (Inclusive method)

Class Interval	Tally marks	Frequency	Relative Frequency
20 - 29	////	4	$\frac{4}{40} = 0.10$
30 - 39	/// ///	8	$\frac{8}{40} = 0.20$
40 - 49	/// ///	10	$\frac{10}{40} = 0.25$
50 - 59	/// ///	8	$\frac{8}{40} = 0.20$
60 - 69	/// /	6	$\frac{6}{40} = 0.15$
70 - 79	////	4	$\frac{4}{40} = 0.10$
Total		$N = 40$	1.00

Table 2: Continuous Frequency Distribution using class boundary (Modified Inclusive Method)

Class Interval	Tally marks	Frequency
19.5 - 29.5	////	4
29.5 - 39.5	/// ///	8
39.5 - 49.5	/// ///	10
49.5 - 59.5	/// ///	8
59.5 - 69.5	/// /	6
69.5 - 79.5	////	4

Table 3: Grouped Frequency Distribution (Exclusive method)

Class Interval	Tally marks	Frequency
20 - 30	////	4
30 - 40	/// ///	8
40 - 50	/// ///	10
50 - 60	/// ///	8
60 - 70	/// /	6
70 - 80	////	4

GRAPHICAL REPRESENTATION

One very simple but effective form of statistical analysis is to present the tabulated data with the help of graphs and diagrams.

BASIC PRINCIPLES OF GRAPHS AND DIAGRAMS

1. Each diagram should have a short, suitable and self-explanatory title.
2. The graphs should be simple and well-defined.
3. The graphs should be completely self-explanatory.
4. The origin, vertical and horizontal scale should be so chosen that a graph does not carry a false impression about the nature of the data
5. Frequency or rate is usually represented on the vertical scale and the variable or method of classification on the horizontal scale.

USES OF GRAPHS

1. It is useful in elucidating the main features of a set of data.
2. It is often valuable in suggesting an appropriate analysis method and explaining the conclusions founded upon the analysis.
3. It can sometimes pinpoint gross errors in statistical records.

DIFFERENCE BETWEEN GRAPH AND DIAGRAM

Both are used as a mode of statistical investigation. However, there are some differences between these two:

1. Graphs are usually drawn on graph paper, whereas diagrams can construct on plain paper.
2. Graphs are helpful to have an idea of the relationship between two correlated variables, whereas diagrams are used to represent the value of an attribute.
3. Graphs are usually used to depict the time series data, while diagram is not used for the same purpose. It is used in depicting categorical and geographical data.
4. Graphs are more precise and accurate than diagrams.
5. Graphs are helpful to indicate further statistical analysis, while diagrams cannot do so.

LIMITATIONS OF GRAPHS AND DIAGRAMS

1. They may be misleading unless drawn and studied with care.
2. It is difficult to draw two or more directional diagrams and also difficult to interpret the such diagram.
3. The conclusions drawn from graphs should normally be regarded as tentative; therefore, the graphs are no substitute for more critical statistical analysis.

TYPES OF GRAPHS AND DIAGRAMS

The important graphs and diagrams which are used for the presentation of statistical data are

(i) Line Diagram
(ii) Bar Diagram
(iii) Pie Diagram
(iv) Pictogram
(v) Histogram
(vi) Frequency Polygon
(vii) Frequency Curve
(viii) Cumulative Frequency Curve or *Ogive*
(ix) Stem-and-leaf Plot
(x) Box and Whisker Plot

AI (Email: atique@stat.jnu.ac.bd)

For quick instance, the line diagram is used for time series data, the bar diagram, pictogram and p chart are used for qualitative data and the rest are used for quantitative data

BAR DIAGRAM

A bar diagram, also known as a bar chart, is a form of presentation in which the frequencies are presented by rectangles usually separated along the horizontal axis and drawn as bars of convenient widths.

EXAMPLE

Let us consider the following data for constructing a bar diagram

Table 1: Health center visit data for constructing a bar diagram

Visit	*No. of person*
Frequently	*49*
Occasionally	*71*
Rarely	*24*
Never	*6*
Total	*150*

The vertical bar diagram constructed from these data is shown in Figure 1:

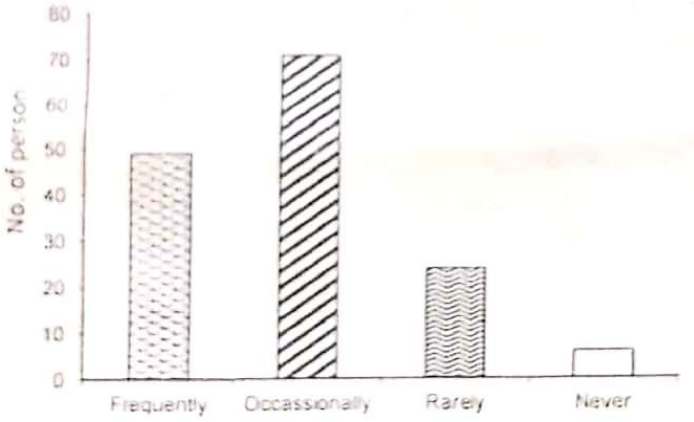


Figure 1: Vertical bar diagram for health center visit data

PIE CHART/DIAGRAM

The pie chart consists of a circle sub-divided into sectors, whose areas are proportional to the various parts into which the whole quantity is divided.

Before that, we form the relative frequency distribution (%) for this purpose and convert the percentage values into angles. As a circle consists of 360° (degree), the whole quantity is equated to 360°. For example, the angle in the degree to the category 'frequently' is arrived at as follows;

$$\frac{49}{150} \times 360 = 117.6^0$$

And for the category occasionally;

$$\frac{71}{150} \times 360 = 170.4^0$$

Other category values are obtained in a similar manner. The necessary computation is shown in Table 2.

Table 2: Health center visit data for constructing a pie diagram

Response	Frequency	Percent Relative Frequency	Angles of The Sector
Frequent	49	32.7	117.6
Occasional	71	47.3	170.4
Rare	24	16.0	57.6
Never	6	4.0	14.4
Total	150	100.0	360.0

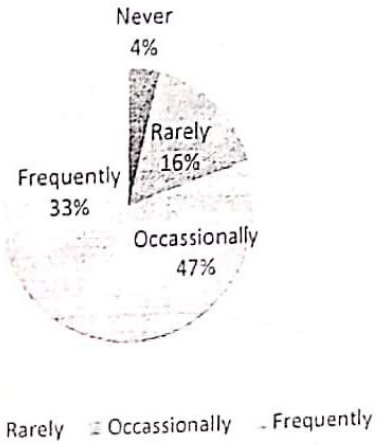


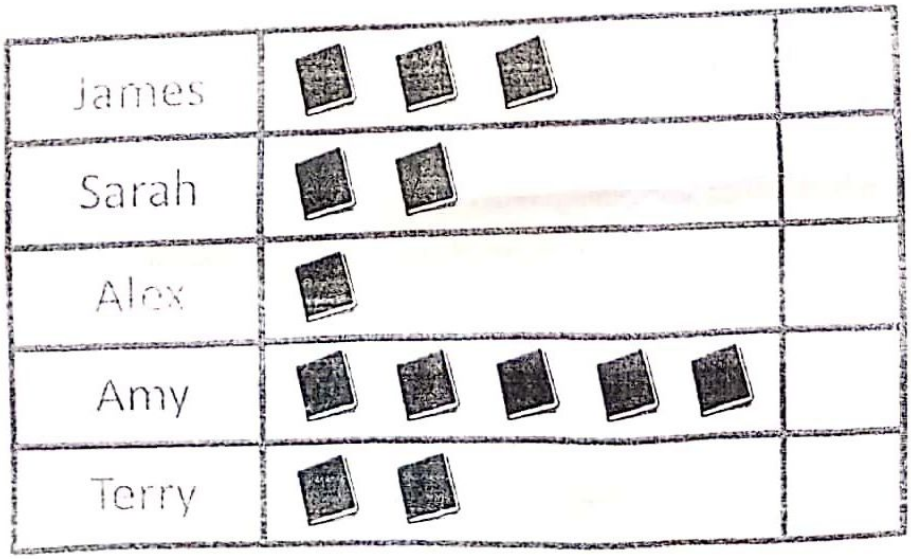
Figure 2: Simple pie diagram for health center visit data.

PICTOGRAM

Also referred to as a picto or a pictograph, a pictogram is a chart or graph used to represent data. Unlike bar graphs or line graphs which have a sole way of representing data, pictograms use pictures with each picture representing a different physical object.

Pictograms are a great way to introduce children to data handling, as they are very visual and generally easy to understand and interpret. Children can simply count the objects to find out how many each option represents.

Example: (a) Who read the most books? (b) Who read the least book?



HISTOGRAM

A histogram is constructed by placing the class boundaries on the horizontal axis and the frequencies on the vertical axis. Each class is shown on the graph by drawing a rectangle whose base is the class boundary and whose height is the corresponding frequency for the class.

Table 3: Frequency Distribution of Age of 20 people

Class Interval	Frequency
20-30	2
30-40	4
40-50	4
50-60	5
60-70	3
70-80	1
80-90	0
90-100	1

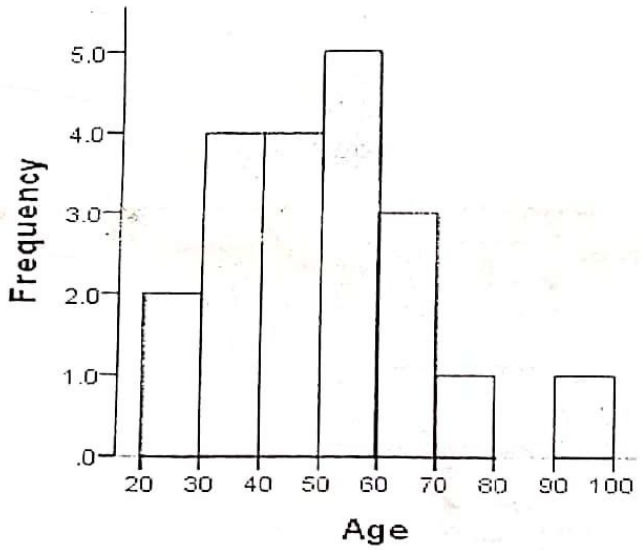


Figure 3: A Histogram presenting the frequency distribution of the age of 20 people.

DIFFERENCE BETWEEN HISTOGRAM AND BAR DIAGRAM

1. A histogram is a two-dimensional diagram, while in a bar diagram, only the length of the diagram is considered (one-dimensional).
2. In the bar diagram, all bars are drawn at equidistance, while bars of the histogram are adjacent.
3. A bar diagram represents the value of a qualitative variable or value against a time period. In contrast, a histogram is used to represent the frequencies of a quantitative variable.
4. In a histogram, the area of a bar depends on the width of the class and the class frequency, while in a bar diagram, the area of a bar depends on the value of a level of the qualitative variable.

STEM AND LEAF PLOT

A Stem and Leaf Plot is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit).

Example

A random sample of 64 people was selected to take the Stanford-Binet Intelligence Test. After each person completed the test, they were assigned an intelligence quotient (IQ) based on their performance on the test. The resulting 64 IQs are as follows:

Table 4: IQ score of 64 people

111	85	83	98	107	101	100	94	101	86
105	122	104	106	90	123	102	107	93	109
141	86	91	88	98	128	93	114	87	116
99	94	94	106	136	102	75	96	78	116
107	106	68	104	91	87	105	97	110	91
107	107	85	117	93	108	91	110	105	99
85	99	99	96						

We could, of course, summarize the data using a histogram. One primary disadvantage of using a histogram to summarize data is that the original data are not preserved in the graph. A **stem-and-leaf plot**, on the other hand, summarizes the data and preserves the data at the same time.

The basic idea behind a stem-and-leaf plot is to divide each data point into a stem and a leaf. We could divide our first data point, 111, for example, into a stem of 11 and a leaf of 1; 85 into a stem of 8 and a leaf of 5; 83 into a stem of 8 and a leaf of 3, and so on.

To create the plot, **we first create a column of numbers containing the ordered stems.** Our IQ data set produces stems 6, 7, 8, 9, 10, 11, 12, 13, and 14.

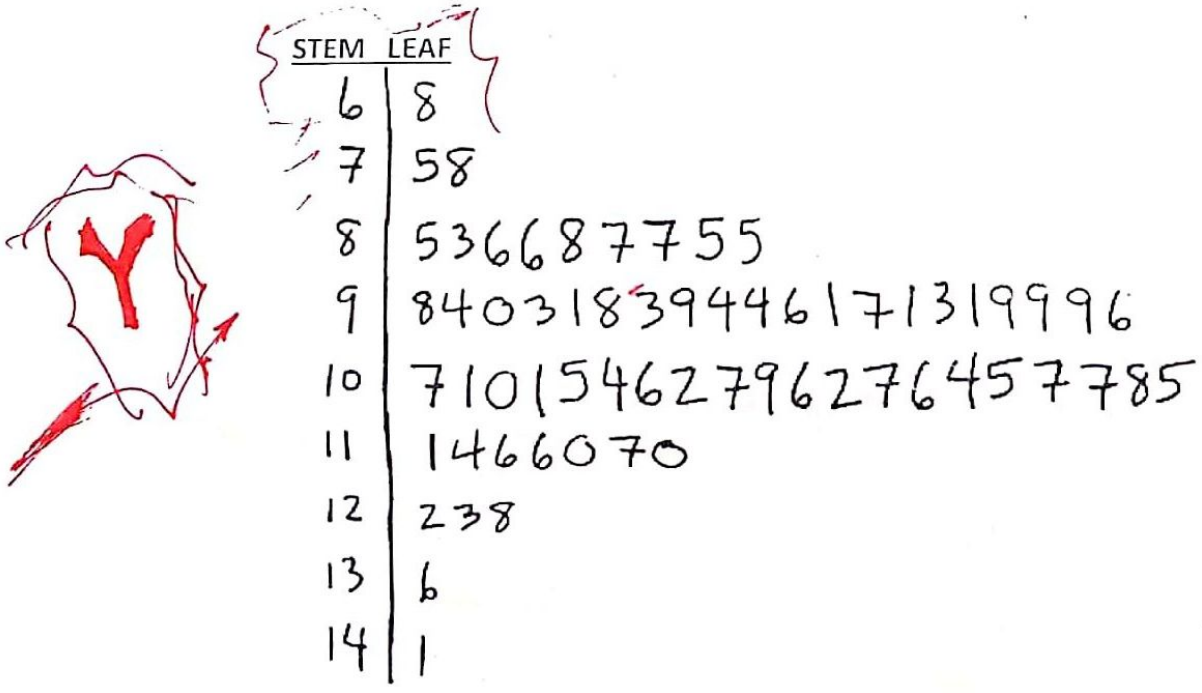


Figure 4: Stem and Leaf plot

FREQUENCY POLYGON

A diagram used to represent a frequency distribution is called a frequency polygon. The mid-values of class intervals are plotted along the X-axis, and corresponding frequencies are plotted along the Y-axis. Later points are joined by straight lines forming with the X-axis is called a frequency polygon.

The frequency polygon for the frequency distribution of expenditure of the elementary school students is shown in the Table in the figure below.

Table 5: Expenditure of the elementary school students

Class interval	Mid-value	Frequency
4.5-8.5	6.5	6
8.5-12.5	10.5	19
12.5-16.5	14.5	23
16.5-20.5	18.5	18
20.5-24.5	22.5	9
24.5-28.5	26.5	3
28.5-32.5	30.5	1
32.5-36.5	34.5	1

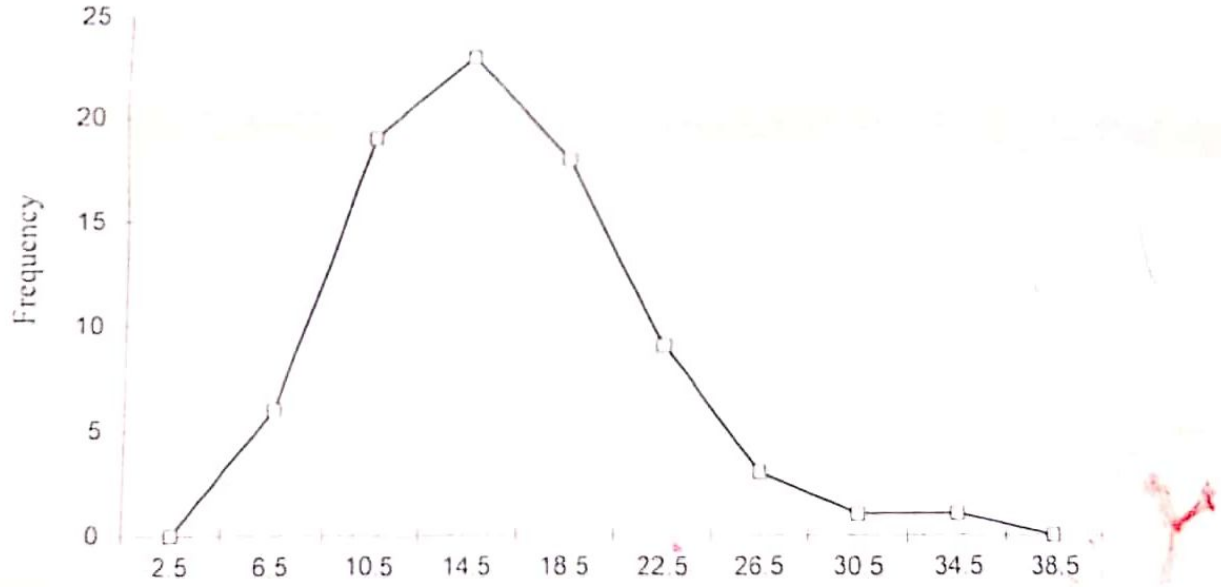


Figure 5: Frequency polygon for the expenditure of the elementary school students

CUMULATIVE FREQUENCY POLYGON or OGIVE

An ogive is based on cumulative frequency distribution. To construct a cumulative frequency distribution, the frequencies are to be cumulated just by summing the class frequencies. Two types of cumulative distributions are used to draw an ogive. *'less than type ogive'* and *'more than type ogive'*

CONSTRUCTING OGIVE

A graph of the cumulative frequency distribution or cumulative relative frequency distribution is called an *ogive*.

To construct a less than type ogive, the upper-class limits (precisely the upper boundaries) are put on the horizontal axis, and cumulative frequencies are shown on the vertical axis. A point is plotted against each upper-class limit at a height corresponding to cumulative frequency at that upper-class limit. One additional point is plotted above the lower-class limit for first class at a zero height. These points are then connected by straight lines. The straight lines allow one to approximate the cumulative frequency between the class limits by interpolating. The resulting graph is a *less than type ogive*.

To construct a *more than type ogive*, a point is plotted against each lower-class limit at a height corresponding to the cumulative frequency at that lower-class limit. As before, an additional point is to be plotted above the upper-class limit for the terminal class at zero height. These points are then connected by straight lines. The resulting graph is a *more than type ogive*.

EXAMPLE

The following Table is constructed from data collected on the life length of 40 rats in years for a laboratory experiment. Display the data by a less than type and a more than type ogive.

Table 6.1: Life length of 40 rats in years

Life length (in years)	Number of Rats
1.45 - 1.95	2
1.95 - 2.45	1
2.45 - 2.95	4
2.95 - 3.45	15
3.45 - 3.95	10
3.95 - 4.45	5
4.45 - 4.95	3
Total	40

This Table is constructed to draw the required ogives, and the resulting ogives are sketched in Figures 6.1 and 6.2

The ogive or cumulative frequency polygon has the advantage of providing a convenient way to estimate the median and the percentiles of a sample. In addition, it has the advantage that the number of items between two values can be readily ascertained. The ogive allows seeing how many observations in a dataset fall in or below a given point on the scale. This is most useful when we have a distribution of scores, and we are interested in finding out how one score compares to the rest of the scores.

Table 6.2: Cumulative frequency distributions for less than and more than type ogives based on the rat life data.

Less than type		More than type	
Age	Cumulative Frequency	Age	Cumulative Frequency
Less than 1.45	0	1.45 or more	40
Less than 1.95	2	1.95 or more	38
Less than 2.45	3	2.45 or more	37
Less than 2.95	7	2.95 or more	33
Less than 3.45	22	3.45 or more	18
Less than 3.95	32	3.95 or more	8
Less than 4.45	37	4.45 or more	3
Less than 4.95	40	4.95 or more	0

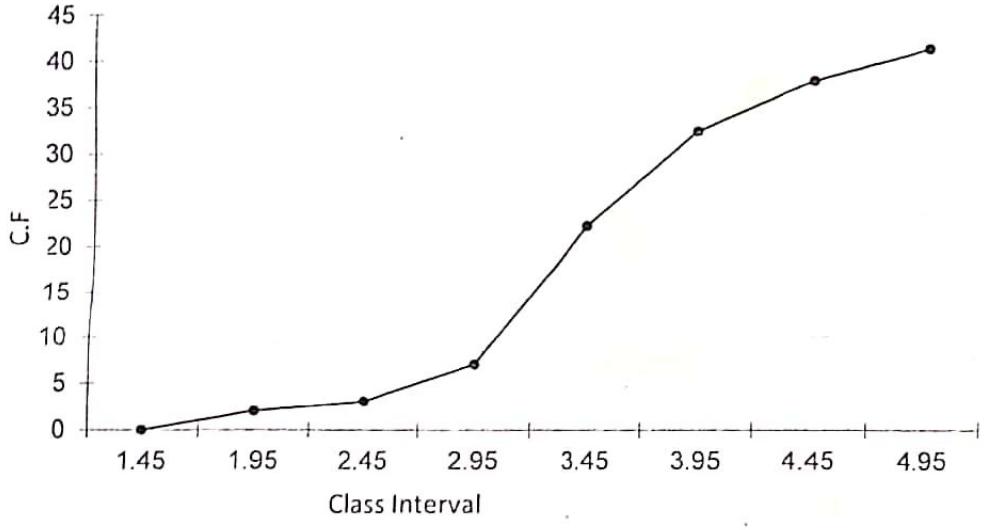


Figure 6.1: Less than type ogive for data in Table 6.2

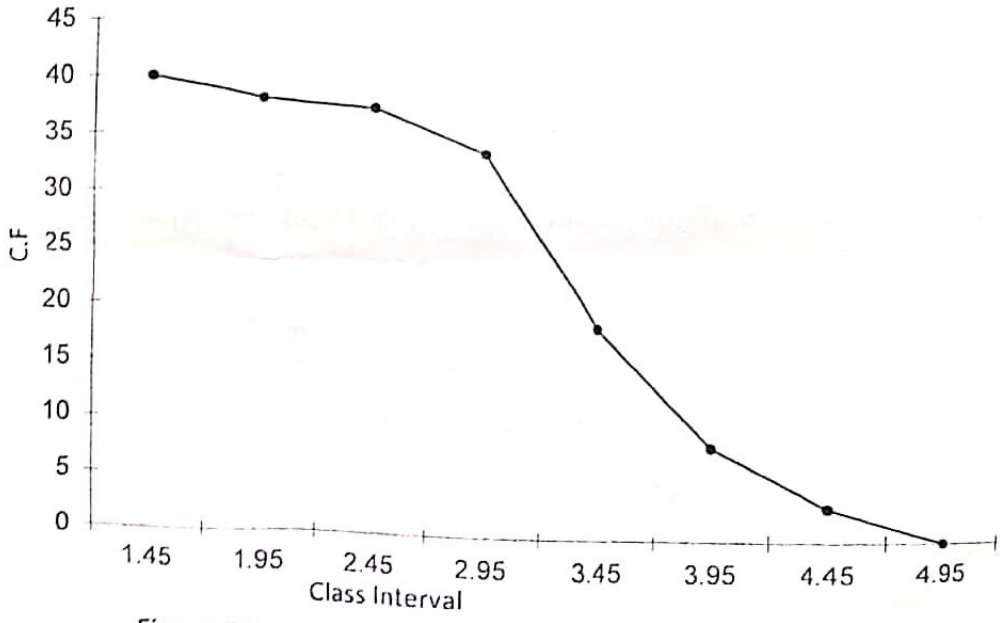


Figure 6.2: More than type ogive for data in Table 6.2

BOX AND WHISKER PLOT

This diagram is used to show the quartiles of the data. A quartile is a measure in which the array is divided into four equal parts, and we have three quartiles, viz. Q_1, Q_2 and Q_3, where,

$$Q_i = Value\ of\ \frac{i\,(n+1)}{4}\ th\ observation$$

The value of Q_1 and Q_3 are plotted, and the box is drawn parallel to the horizontal axis, where values of the variable are shown in the horizontal axis. The box is divided at the point of value of Q_2. The two ends of the box are extended to the lower and upper limits of the values. These extended lines are called **Whisker**.

Briefly, A box-and-whisker plot or boxplot is a diagram based on the five-number summary of a dataset.

QUARTILES

Quartiles are those values in which the total frequency is divided into four equal parts. There are three quartiles, i.e., 1st quartile (Q_1), 2nd quartile (Q_2), and the third quartile (Q_3). Second quartile Q_2 is equal to the median. The quartiles are calculated by

$$Q_i = Value\ of\ \frac{i\,(n+1)}{4}\ th\ observation$$

EXAMPLE: Calculate the 1st, 2nd, and 3rd quartiles from the following data 10, 50, 30, 20, 10, 20, 70, 30.

SOLUTION: Arranging observations in ascending order, we get:

10, 10, 20, 20, 30, 30, 50, 70

Here, $n = 8$, minimum value = 10, maximum value = 70. So, the 1st quartile is

$$Q_1 = \frac{1 \times (8+1)}{4}\ th\ value\ of\ the\ obervation$$

$$\Rightarrow Q_1 = 2.25th\ value\ of\ the\ observation$$

$$\Rightarrow Q_1 = 2nd\ observation + 0.25 \times (3rd - 2nd\ observation)$$

$$\Rightarrow Q_1 = 10 + 0.25 \times (20 - 10) = 10 + 2.5 = 12.5$$

2nd quartile or median is

$$Q_2 = \frac{2 \times (8+1)}{4}\ th\ value\ of\ the\ obervation$$

$$\Rightarrow Q_2 = 4.5th\ value\ of\ the\ observation$$

$$\Rightarrow Q_2 = 4th\ observation + 0.5 \times (5th - 4th\ observation)$$

$$\Rightarrow Q_2 = 20 + 0.5 \times (30 - 20) = 20 + 5 = 25$$

3rd quartile is

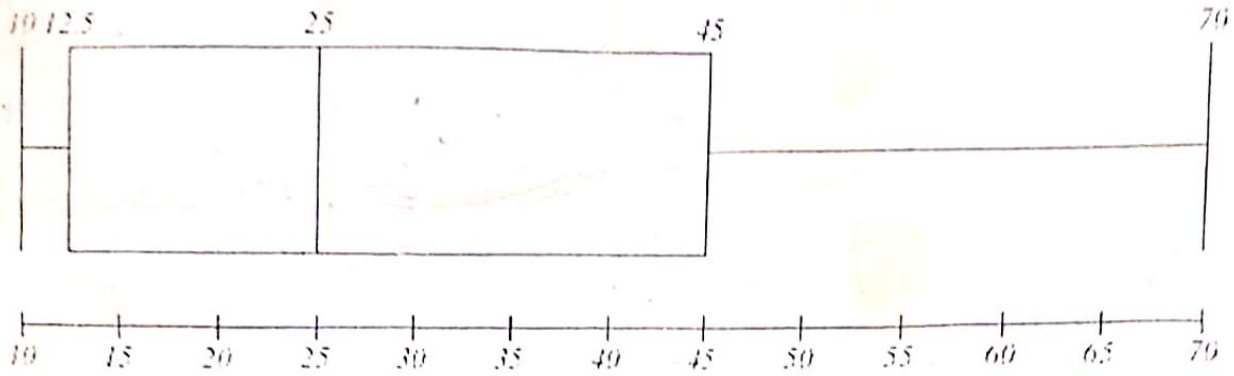
$$Q_3 = \frac{3 \times (8+1)}{4}\ th\ value\ of\ the\ obervation$$

$$\Rightarrow Q_3 = 6.75th\ value\ of\ the\ observation$$

$$\Rightarrow Q_3 = 6th\ observation + 0.75 \times (7th - 6th\ observation)$$

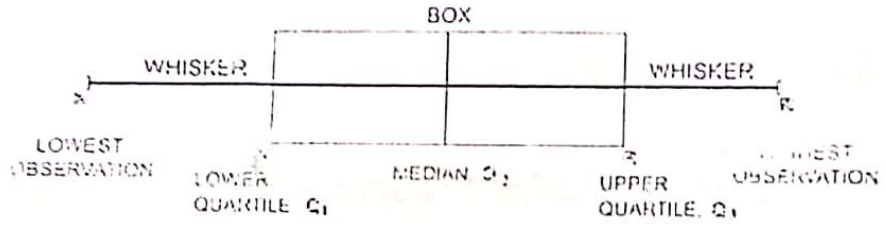
$$\Rightarrow Q_3 = 30 + 0.75 \times (50 - 30) = 30 + 15 = 45$$

The box-and-whisker plot is given by



FIVE-NUMBER SUMMARY

The five-number summary of a data set consists of the five numbers determined by computing the **minimum**, Q_1, **median**, Q_3, **and maximum** value of the data set.



EXAMPLE

Mr. X works at a computer store. He also recorded the number of sales he made each month. In the past 11 months, he sold the following numbers of computers:
51, 17, 25, 39, 7, 49, 62, 41, 20, 43, 13.

(i) Give a five-number summary of Mr. X's sales.
(ii) Make a box and whisker plot for Mr. X's sales.

Location or Position of a distribution

The location or position of a frequency distribution is the value of the variable around which most of the values in the distribution tend to cluster.

The measure of Central Tendency

The measure which usually reflects the complete dataset and falls in the centre of the array is known as the measure of central tendency since it tends to lie in the centre.

Example: An average of a set of values of a variable which is typical or representative of the whole data.

Characteristics of an Ideal Measure of Central Tendency

1. It should be rigidly defined so that its value is stable and presents the whole data set.
2. It should be easy to understand and calculate with accuracy, even by a non-mathematical person.
3. It should be based on all the observations of the data set.
4. It should be so defined that the measure is amenable to mathematical treatment in further analysis.
5. It should be affected as small as possible by sampling fluctuation.
6. It should not be affected by extreme observations.

- **The following are the main measures of Central Tendency**
  1) Mean
  2) Median
  3) Mode

Mean is of three types:

i). Arithmetic Mean
ii). Geometric Mean
iii). Harmonic Mean

(i) Arithmetic Mean (AM): Arithmetic mean of a set of observations is their sum divided by the number of observations, e.g., the AM, $\bar{x}$, of n observations $x_1, x_2, \ldots, x_n$ is given by

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots\ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

In the case of frequency distribution, where f_i is the frequency of the variable x, i.e.

Value (x_i)	Frequency (f_i)	$f_i x_i$
x_1	f_1	$f_1 x_1$
x_2	f_2	$f_2 x_2$
⋮	⋮	⋮
x_k	f_k	$f_k x_k$

$$\bar{x} = \frac{f_1 x_1 + f_2 x_2 + \cdots\ldots\ldots + f_k x_k}{f_1 + f_2 + \cdots\ldots + f_k} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{\sum_{i=1}^{k} f_i x_i}{N}$$

In the case of grouped or continuous frequency distribution, x is taken as the mid-value of the corresponding class.

Example: Find the arithmetic mean from the following frequency distribution

Class Interval	Frequency
11-13	3
13-15	4
15-17	5
17-19	10
19-21	6
21-23	4
23-25	3

Solution:

Class Interval	Mid-value (x_i)	Frequency (f_i)	$f_i x_i$
11-13	12	3	36
13-15	14	4	56
15-17	16	5	80
17-19	18	10	180
19-21	20	6	120
21-23	22	4	88
23-25	24	3	72
		$\sum f_i = 35$	$\sum f_i x_i = 632$

$$\therefore\ AM = \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{\sum_{i=1}^{k} f_i} = \frac{632}{35} = 18.06.$$

Properties of Arithmetic Mean:

(i) The algebraic sum of the deviations of the observations from the arithmetic mean is zero. i.e. $\sum_{i=1}^{k} f_i(x_i - \bar{x}) = 0$.

Proof:

$$\sum_{i=1}^{k} f_i(x_i - \bar{x}) = \sum_{i=1}^{k} f_i x_i - \bar{x}\sum_{i=1}^{k} f_i = N\bar{x} - N\bar{x} = 0$$

$$= 0.$$

$$\text{Where,}\quad \bar{x} = \frac{\sum_{i=1}^{k} f_i x_i}{N} \Rightarrow N\bar{x} = \sum_{i=1}^{k} f_i x_i$$

(ii) Sum of squares of the deviations of a set of observations is minimum when the deviations are taken about the AM. i.e. $\sum_{i=1}^{k} f_i(x_i - A)^2 > \sum_{i=1}^{k} f_i(x_i - \bar{x})^2$.

Proof: Let $\bar{x}$ be the arithmetic mean and A is an arbitrary value.

$$\sum_{i=1}^{k} f_i(x_i - A)^2 = \sum_{i=1}^{k} f_i\{(x_i - \bar{x}) + (\bar{x} - A)\}^2$$

$$= \sum_{i=1}^{k} f_i(x_i - \bar{x})^2 + 2(\bar{x} - A)\sum_{i=1}^{k} f_i(x_i - \bar{x}) + (\bar{x} - A)2\sum_{i=1}^{k} f_i$$

$$= \sum_{i=1}^{k} f_i(x_i - \bar{x})^2 + N(\bar{x} - A)^2 \quad \text{Since } \sum_{i=1}^{k} f_i(x_i - \bar{x}) = 0.$$

$$Therefore, \sum_{i=1}^{k} f_i(x_i - A)^2 > \sum_{i-1}^{k} f_i(x_i - \bar{x})^2.$$

(iii) **Arithmetic mean depends on origin and scale.**

Proof: Let $x_1, x_2, \ldots, x_k$ be the mid values of a frequency distribution with equal class interval and the corresponding frequencies be $f_1, f_2, \ldots, f_k$ respectively. Then the mean of the distribution is

$$\bar{X} = \frac{1}{N}\sum_{i=1}^{k} f_i x_i, \quad where\ N = \sum_{i=1}^{k} f_i$$

$$Let\ u_i = \frac{x_i - A}{h}, \qquad (i = 1,2,3,\ldots,k)$$

where, u_i is a transformed variable, A is the provisional mean taken to be equal to one of the mid-values, and h is the size of the class interval.

The above relation can be rewritten as follows:

$$x_i = A + hu_i$$

$$or, \quad f_i x_i = Af_i + hf_i u_i$$

$$or, \quad \sum_{i=1}^{k} f_i x_i = \sum_{i=1}^{k} Af_i + h\sum_{i=1}^{k} f_i u_i \quad [Summing\ over\ both\ sides\ over\ i\ from\ 1\ to\ k]$$

$$or, \quad \frac{1}{N}\sum_{i=1}^{k} f_i x_i = \frac{A}{N}\sum_{i=1}^{k} f_i + \frac{h}{N}\sum_{i=1}^{k} f_i u_i$$

$$or, \quad \bar{X} = \frac{A}{N}.N + h\bar{u}$$

$$or, \quad \bar{X} = A + h\bar{u}$$

Thus, arithmetic mean $\bar{X}$ depends on change of origin A and scale h.

Merits:

1. It is rigidly defined.
2. It is easy to understand and easy to calculate.
3. It is based upon all the observations.
4. It is amenable to further algebraic treatment.
5. Of all the averages, AM affected least by sampling fluctuation.

Demerits:

1. It cannot be determined by inspection nor can it be located graphically.
2. AM cannot be used if we are dealing with qualitative characteristics, such as intelligence, honesty, beauty etc.

3. AM cannot obtain if a single observation is missing or lost unless we drop it out and compute the arithmetic mean of the remaining values.
4. AM is affected very much by extreme values.
5. It does not provide exact value for open-end classes.

Weighted Arithmetic Mean:

The weighted mean is defined as an average computed by giving different weights to some of the individual values. When all the weights are equal, then the weighted mean is similar to the arithmetic mean.

Sometimes we associate with the numbers $x_1, x_2, \dots, x_n$ certain weighting factors or weights $w_1, w_2, \dots\dots\dots, w_n$ depending on the significance or importance attached to the numbers. In this case,

$$\bar{x}_w = \frac{w_1x_1 + w_2x_2 + \dots\dots\dots + w_nx_n}{w_1 + w_2 + \dots\dots + w_n} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i}$$

is called weighted arithmetic mean.

Example: If a final exam in a course is weighted 3 times as much as a quiz and a student has a final exam grade of 85 and quiz grades 70 and 90, the mean grade is

$$\bar{x}_w = \frac{1 \times 70 + 1 \times 90 + 3 \times 85}{1 + 1 + 3} = \frac{415}{5} = 83.$$

Problem: Find the simple and weighted arithmetic mean of the first n natural numbers, the weights being the corresponding numbers.

Solution: The first n natural numbers are $1, 2, 3, \dots\dots, n.\ (n > 0)$.

Let us construct the following table:

x_i	w_i	w_ix_i
1	1	1^2
2	2	2^2
3	3	3^2
⋮	⋮	⋮
n	n	n^2

We know that, $1 + 2 + 3 + \dots\dots + n = \frac{n(n+1)}{2}$ and

$$1^2 + 2^2 + 3^2 + \dots\dots + n^2 = \frac{n(n+1)(2n+1)}{6}$$

Therefore, Simple arithmetic mean (AM) is,

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\frac{n(n+1)}{2}}{n} = \frac{n(n+1)}{2} \cdot \frac{1}{n} = \frac{n+1}{2}$$

And weighted AM is

$$\overline{X_w} = \frac{\sum_{i=1}^{n} w_ix_i}{\sum_{i=1}^{n} w_i} = \frac{\frac{n(n+1)(2n+1)}{6}}{\frac{n(n+1)}{2}} = \frac{n(n+1)(2n+1)}{6} \times \frac{2}{n(n+1)} = \frac{2n+1}{3}$$

$$Therefore, Simple\ AM = \frac{n+1}{2}\ and\ weighted\ AM = \frac{2n+1}{3}.$$

Short-cut Method of Computing Arithmetic Mean:

Let h be the width of class interval and a be an arbitrary mid-point. We then define U as follows:

$$u_i = \frac{x_i - A}{h}$$

$$\Rightarrow x_i = A + hu_i$$

$$\Rightarrow \bar{x} = A + h\bar{u}$$

$$where, \bar{u} = \frac{\sum_{i=1}^{k} f_i U_i}{\sum_{i=1}^{k} f_i}$$

Consider the following example:

Class Interval	Mid-value (x_i)	Frequency (f_i)	$u_i = \frac{x_i - a}{h}$	$f_i u_i$
48.5 – 53.5	51	2	-4	-8
53.5 – 58.5	56	2	-3	-6
58.5 – 63.5	61	3	-2	-6
63.5 – 68.5	66	5	-1	-5
68.5 – 73.5	71	5	0	0
73.5 – 78.5	76	5	1	5
78.5 – 83.5	81	5	2	10
83.5 – 88.5	86	7	3	21
88.5 – 93.5	91	10	4	40
93.5 – 98.5	96	6	5	30
		$\sum_{i=1}^{10} f_i = 50$		$\sum_{i=1}^{10} f_i u_i = 81$

$$Here, A = 71, h = 5 \text{ and } u_i = \frac{x_i - 71}{5}$$

$$Thus, \bar{u} = \frac{\sum_{i=1}^{k} f_i u_i}{\sum_{i=1}^{k} f_i} = \frac{81}{50} = 1.62$$

$$\therefore \bar{x} = A + h\bar{u} = 71 + 5 \times 1.62 = 79.1$$

(ii) Geometric Mean: Geometric Mean of a set of n non-zero positive observations is the nth root of their product.

Let $x_1, x_2, \ldots, x_n$ be n non-zero positive observations of a series of data. Then the GM is,

$$GM = G = (x_1 . x_2 . \ldots . x_n)^{\frac{1}{n}}$$

The calculation may sometimes be simplified by taking logarithm,

$$LogG = \frac{1}{n}\sum_{i=1}^{n} logx_i$$

$$\Rightarrow G = Antilog\left[\frac{1}{n}\sum_{i=1}^{n} logx_i\right]$$

In case of frequency distribution, GM is given by,

$$G = \left[x_1^{f_1} x_1^{f_2} \ldots x_k^{f_k}\right]^{\frac{1}{N}}, where, N = \sum_{i=1}^{k} f_i$$

$$\Rightarrow logG = \frac{1}{N}\sum_{i=1}^{k} f_i log\, x_i$$

$$\Rightarrow G = Antilog\left[\frac{1}{N}\sum_{i=1}^{k} f_i log\, x_i\right]$$

Example 1: Suppose the sales of a departmental store had increased from taka 2 lac in 1990 to taka 4 lac in 1991 and taka 6 lac in1992. What is the average rate of increase in sales per year?

Solution: *The increase is* $(\frac{6-2}{2}) \times 100 = 200$ *percent over the 2-year period 1990-92. This is obviously wrong since the average rate of increase per year was much less.*

The appropriate average in this instance is the geometric mean. Since the sales in 1991 was twice as high as in1990, and the sales in1992 was 1.5 times as high as in 1991, the GM of these two values is $G = \sqrt{2 \times 1.5} = \sqrt{3} = 1.7325$. Thus, the average rate of increase in sales per year is $1.7325 - 1 = 0.7325\ i.e.\ 73.25\%$.

Example 2: In Growth Rates

The geometric mean is used in finance to calculate average growth rates and is referred to as the *compounded annual growth rate*. Consider a stock that grows by 10% in year one, declines by 20% in year two, and then grows by 30% in year three. What is the growth rate of the company?

Solution: The geometric mean of the growth rate is calculated as follows:

$$GM = \left((1 + 0.1) * (1 - 0.2) * (1 + 0.3)\right)^{\frac{1}{3}} - 1 = 0.046\ or\ 4.6\%\ annually.$$

Merits:

1. It is rigidly defined.
2. It is based upon all the observations.
3. It is amenable to further algebraic treatment.
4. It is not affected much by sampling fluctuations.
5. It gives comparatively more weight to small items.

Demerits:

1. Because of its abstract mathematical character GM is not easy to understand and to calculate for a non-mathematics student.
2. If any one of the observations is zero, GM becomes zero and if any one observation is negative, GM becomes imaginary regardless of the magnitude of the other items.

Uses: GM is used –

1. To calculate average of ratios and percentages.
2. To calculate the growth rate of population and average rate of increase or decrease in economics activities.
3. To calculate index number.

(iii) Harmonic Mean: The Harmonic Mean of a set of n non-zero observations $x_1, x_2, \dots, x_n$ in a series is the reciprocal to the arithmetic mean of the reciprocals. That is,

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$$

In case of frequency distribution,

$$HM = \frac{N}{\sum_{i=1}^{k} \frac{f_i}{x_i}}, \quad where \sum f_i = N$$

Example 1: Suppose an aero plane flies around a square with 100 miles long side, taking the first side at 100m/h, the second side at 200m/h, the third side at 300 m/h and fourth side at 400 m/h. What is the average speed of the plane in its flight around the square?

Solution: Here the HM is the appropriate measure for the average speed.

$$HM = \frac{4}{\frac{1}{100} + \frac{1}{200} + \frac{1}{300} + \frac{1}{400}} = \frac{4 \times 1200}{25} = 192\ m/h$$

Example 2: A train moves first 50 km at a sped of 60 km/hour, second 50 km at a speed of 75 km/hour, third 50 km at a speed of 65 km/hour and fourth 50 km at a speed of 80 km/hour. What is the average speed of the train throughout the journey?

Solution: $Given\ x_1 = 60, x_2 = 75, x_3 = 65, and\ x_4 = 80.$

Since the train covers same distance at each step, the distance can be ignored in calculating the average speed.

The average speed (harmonic mean) is given by

$$HM = \frac{n}{\sum_{i=1}^{n}\frac{1}{x_i}} = \frac{4}{\frac{1}{60}+\frac{1}{75}+\frac{1}{65}+\frac{1}{80}} = 69.10\ km/hour.$$

Example 3: A plane moves 1st 800 km at a speed of 600 km/hour, second 400 km at a speed of 800 km/hour and last 200 km at a speed of 500 km/hour. Find the average speed of the plane.

Solution: Let the speeds be $x_1 = 600$, $x_2 = 800$, and $x_3 = 500$ and the distances covered by the plane be $d_1 = 800$, $d_2 = 400$, and $d_3 = 200$. The time taken to cover the distances are

$$t_1 = \frac{d_1}{x_1} = \frac{800}{600}, t_2 = \frac{d_2}{x_2} = \frac{400}{800}, and\ t_3 = \frac{d_3}{x_3} = \frac{200}{500}.$$

The average speed (i.e., HM) of the plane is

$$HM = \frac{Total\ distance}{Total\ time} = \frac{800+400+200}{\frac{800}{600}+\frac{400}{800}+\frac{200}{500}} = 626.86\ km/hour.$$

The average speed is known as weighted harmonic mean, where weights are the distances covered.

Merits:

1. It is rigidly defined.
2. It is based upon all the observations.
3. It is amenable to further algebraic treatment.
4. It is not affected much by sampling fluctuations.
5. It gives greater importance to small items and is useful only when small items have to give very high weightage.

Demerits:

1. It is not easy to understand and is difficult to calculate.
2. It cannot be calculated if any observation is zero.
3. It is impossible to calculate for open-end class interval.

Uses: HM is used –

1. To calculate average speed of any vehicle.
2. To calculate the average growth rate and average profit in any business.

Quadratic Mean or Root Mean Square (QM or RMS): If we have a set of n observations represented by $x_1, x_2, \ldots, x_n$, then the QM of these values is given by

$$Q = \sqrt{\frac{x_1^2 + x_2^2 + \ldots\ldots + x_n^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

This mean is seen to have applications in physical and engineering sciences. In power distribution systems, e.g. voltages and electricity are usually referred to in terms of their root mean square value.

2) Median. The median is defined as the middle most observation when the observations are arranged in order of magnitude (ascending or descending).

For ungrouped data, when n is odd, the median is the middle most observation, i.e., $\left(\frac{n+1}{2}\right)^{th}$ observation is the series.

For example – *Let a series be* $5, 10, 7, 3, 2, where, n = 5$

Therefore, ascending order of this series be $2, 3, 5, 7, 10$

Hence, the median of the series is 5.

Again, when n is even, the median will be the arithmetic mean of $\frac{n}{2}^{th}$ and $\left(\frac{n+1}{2}\right)^{th}$ observations in the series.

e.g. - Let us consider the values 11,3,9,5,7,12,15,18; $n = 8$

Ascending order: 3,5,7,9,11,13,15,18

So, the median is $\frac{9+11}{2} = 10$, i.e., $[\frac{8}{2}th = 4th \;\&\; \left(\frac{8}{2}+1\right)th = 5th]$

For grouped frequency distribution the median is given by

$$Me = L + \left(\frac{\frac{N}{2} - F_c}{f}\right) \times h$$

where,

L = the lower limit of the median class (median class is that class which contain $\frac{N}{2}th$ observation of the series)

N = total number of observations,

F_c = cumulative frequency of the class preceding the median class,

f = frequency of the median class,

h = Length of the median class.

Example. Consider the following frequency distribution

Class Boundaries	Frequency (f)	Cumulative Frequency (C.F.)
29.5-39.5	10	10
39.5-49.5	30	40
49.5-59.5	70	110
59.5-69.5	200	310
69.5-79.5	150	460
79.5-89.5	120	580
89.5-99.5	20	600

$$Me = L + \left(\frac{\frac{N}{2} - F_c}{f}\right) \times h$$

Here, $\frac{N}{2} = \frac{600}{2} = 300$

Therefore, the median class is 59.5-69.5. $L = 59.5$, $F_c = 110$, $f = 200$, $h = 10$

Therefore, $Me = 59.5 + \frac{300-110}{200} \times 10 = 59.5 + 9.5 = 69.$

Merits.

1. It is rigidly defined.
2. It is easy to understand and easy to calculate. In some cases, it can be located merely by inspection.
3. It is not at all affected by extreme values.
4. It can be calculated for distributions with open-end classes.

Demerits.

1. In case of even number of observations median cannot be determined exactly.
2. It is not based on all the observations.
3. It is not amenable to algebraic treatment.
4. As compared with AM, it is affected much by sampling fluctuations.

Uses.

1. Median is the only average to be used while dealing with qualitative data which cannot be measured quantitatively but still can be arranged in ascending or descending order of magnitude, e.g., to find the average intelligence or average honesty among a group of people.
2. It is to be used for determining the typical value in problems concerning wages, distribution of wealth etc.

5. Mode. The mode is the value of the variable that occurs most frequently, i.e., for which the frequency is maximum. It is the most fashionable value of the variable. The mode is often denoted by M_o and is frequently referred to as the modal value.

For example – The number of children in 10 families be as follows:

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

Then the mode of this series is 3 because this value occurs most frequently in the series.

For grouped frequency distribution, the mode can be obtained from the formula

$$M_0 = L + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

$$= L + \frac{f_1 - f_0}{(f_1 - f_0) + (f_2 - f_0)} \times h$$

where,

L = lower limit of the modal class.

Δ_1 = the difference between the frequency of the modal class and pre-modal class (f_1-f_0).

Δ_2 = the difference between the frequency of the modal class and post-modal class (f_1-f_2).

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class following the modal class,

h = length of the modal class.

Example. Consider the following frequency distribution

Class Boundaries	Frequency (f)	Cumulative Frequency (C.F.)
29.5-39.5	10	10
39.5-49.5	30	40
49.5-59.5	70	110
59.5-69.5	200	310
69.5-79.5	150	460
79.5-89.5	120	580
89.5-99.5	20	600

Here, the Modal class is 59.5-69.5.

$L_1 = 59.5$, $\Delta_1 = 200\text{-}70 = 130$, $\Delta_2 = 200\text{-}150 = 50$, $h = 10$

Therefore, $M_0 = 59.5 + \frac{130}{130+50} \times 10 = 59.5 + 7.22 = 66.72$.

Merits.

1. Mode is readily comprehensible and easy to calculate. Like median, in some cases it can be located merely by inspection.
2. It is not at all affected by extreme values.
3. Mode can be conveniently located even if the frequency distribution has class intervals of unequal magnitude provided the modal class and the classes preceding and succeeding it are of the same magnitude. Open-end classes also do not pose any problem in the location of mode.

Demerits.

1. Mode is ill-defined. It is not always possible to find a clearly defined mode.
2. It is not based on all the observations.
3. It is not capable of further mathematical treatment.
4. As compared with AM, it is affected to greater extent by sampling fluctuations.

Uses.

Mode is the average to be used to find the ideal size, e.g. in business forecasting, in manufacturing of ready-made garments, shoes, etc.

Other Measures of Location

Quantiles: The quartiles, deciles and percentiles are collectively known as quantiles. Quantiles are those values in a series, which divide the total frequency into number of equal parts when the series is arranged in order of magnitude.

Quartiles: Quartiles are those values, which divide the total frequency into four equal parts. There are three quartiles, i.e., 1st quartile (Q_1), 2nd quartile (Q_2) and the third quartile (Q_3). Second quartile Q_2 is equal to the median.

For group frequency distribution, the quartiles are given by

$$Q_i = L_i + \frac{\frac{iN}{4} - F_i}{f_i} \times h, \qquad i = 1, 2, 3.$$

Where, L_i = lower limit of the ith quartile class

N= total number of observations,

F_i = cumulative frequency of the class preceding the quartile class,

f_i = frequency of the quartile class,

h= Length of the quartile class.

Deciles. Deciles are those values which divide the total frequency into 10 equal parts. There are 9 deciles. These are $D_1, D_2, \ldots, D_9$. The 5th decile or D_5 is the second quartile or median ($D_5 = Q_2 = M_e$).

For grouped frequency distribution, the deciles are given by

$$D_j = L_j + \frac{\frac{j \times N}{10} - F_j}{f_j} \times h, \qquad j = 1, 2, \ldots, 9$$

where L_j = lower limit of the jth decile class.

N = total number of observation (frequency);

F_j = Cumulative frequency of the preceding jth decile class;

f_j = frequency of the jth decile class, and

h = length of class interval of the jth decile class.

Percentiles. Percentiles are those values which divided the total frequency into 100 equal parts. Thus, we have 99 percentiles. These are $P1 \le P2 \le P3 \le \cdots \le P50 \le \cdots \le P99$. The median is the 50th percentile P_{50} as well as 5th decile (D_5) and 2nd quartile (Q_2).

For grouped frequency distribution, the percentiles are given by,

$$P_k = L_k + \frac{\frac{k \times N}{100} - F_k}{f_k} \times h, \qquad k = 1, 2, \ldots, 99$$

where L_k = lower limit of the kth percentile class.

N = total number of observation (frequency);

F_k = Cumulative frequency of the preceding kth percentile class;

f_k = frequency of the kth percentile class, and

h = length of class interval of the kth percentile class.

Example: Find (a) 1st and 3rd quartile, (b) 7th decile and (c) 62nd percentile from the frequency distribution given below:

Class Interval	Frequency	Cumulative frequency
11-13	3	3
13-15	4	7
15-17	5	12
17-19	10	22
19-21	6	28
21-23	4	32
23-25	3	35

(a) (15-17) is the 1st quartile class because $\frac{N}{4}th = \frac{35}{4}th = 8.75^{th}$ observation lies in that class.

$$\text{Hence, } Q_1 = 15 + \frac{8.75-7}{5} \times 2 = 15 + 0.70 = 15.70$$

$\frac{3\times N}{4} = \frac{3\times 35}{4} = 26.25th$ observation falls in the class (19-21).

$$\text{Hence, } Q_3 = 19 + \frac{26.25-22}{6} \times 2 = 20.42.$$

(b) 7th decile class is (19-21), because $\frac{7\times N}{10} = \frac{7\times 35}{10} = 24.5$

$$\text{Hence, } D_7 = 19 + \frac{24.5-22}{6} \times 2 = 19.83.$$

(c) 62nd percentile class is (17-19), because $\frac{62\times N}{100} = \frac{62\times 35}{100} = 21.7$

$$\text{Hence, } P_{62} = 17 + \frac{21.7-12}{10} \times 2 = 18.94.$$

Box and Whisker plot

This diagram is used to show the quartiles of the data. Quartile is the measure which divides the array into 4 equal parts, and we have 3 quartiles, viz. Q_1, Q_2 and Q_3
Where,

$$Q_i = Value\ of\ \frac{i\,(n+1)}{4}th\ observation, if\ n\ is\ odd$$

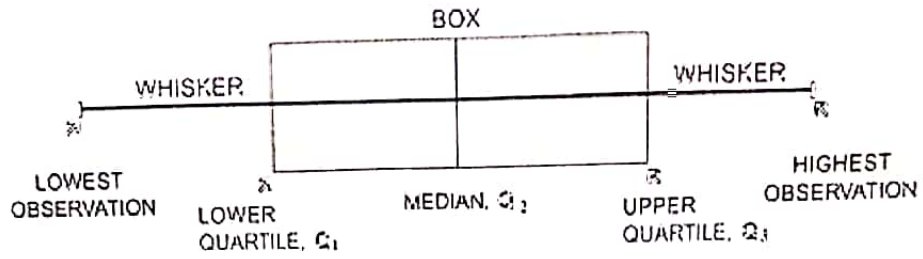
$$= Value\ of\ \frac{1}{2}\left[\frac{in}{4}th + \left(\frac{in}{4}+1\right)th\right] observation, if\ n\ is\ even$$

The value of Q_1 and Q_3 are plotted and box is drawn parallel to the horizontal axis, where values of the variable are shown in the horizontal axis. The box is divided at the point of value of Q_2. The two ends of the box are extended to the lower and upper limit of the values. These extended lines are called **Whisker**.

Briefly, A box-and-whisker plot or boxplot is a diagram based on the five-number summary of a data set.

Five-number summary

The five-number summary of a data set consists of the five numbers determined by computing the **minimum, Q_1, median, Q_3 and maximum** value of the data set.



Example

Mr. X works at a computer store. He also recorded the number of sales he made each month. In the past 11 months, he sold the following numbers of computers:
51, 17, 25, 39, 7, 49, 62, 41, 20, 43, 13.

(i) Give a five-number summary of Mr. X sales.
(ii) Make a box and whisker plot for Mr. X sales.

Theorem 1: For two non-zero positive observations prove that $AM \times HM = GM^2$.

Proof: Let the two observations be X_1 and X_2 X_1, Then, by definition

$$AM = \frac{X_1 + X_2}{2},\ GM = (X_1.X_2)^{\frac{1}{2}} \text{ and } HM = \frac{1}{\frac{1}{2}\left(\frac{1}{X_1} + \frac{1}{X_2}\right)} = \frac{2X_1X_2}{X_1 + X_2}$$

$$\text{Therefore, } AM \times HM = \frac{X_1 + X_2}{2} \cdot \frac{2X_1X_2}{X_1 + X_2} = X_1X_2$$

$$= \left\{(X_1X_2)^{\frac{1}{2}}\right\}^2 = GM^2$$

Hence, $AM \times HM = GM^2$.

Theorem 2: For a set of n non-zero positive values $x_1, x_2, \ldots, x_n$, prove that $AM \geq GM \geq HM$.

Proof: By Definition, $AM = A = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$\text{and } GM = G = (x_1.x_2.....x_n)^{\frac{1}{n}}$$

Taking logarithm of G

$$log\ G = \frac{1}{n} log(x_1.x_2.....x_n) = \frac{1}{n}\sum_{i=1}^{n} log x_i$$

$$= \frac{1}{n}\sum_{i=1}^{n} log\left[\frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} x_i}{n} + x_i\right]$$

$$= \frac{1}{n}\sum_{i=1}^{n} log\,(A - A + x_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} log\, A\left(1 - \frac{A - x_i}{A}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n} log\, \Lambda + \frac{1}{n}\sum_{i=1}^{n} log\left(1 - \frac{\Lambda\text{-}x_i}{\Lambda}\right)$$

$$= \frac{1}{n}.n\, log\, \Lambda + \frac{1}{n}\sum\left\{-\frac{\Lambda - x_i}{\Lambda} - \frac{1}{2}\left(\frac{\Lambda - x_i}{\Lambda}\right)^2 - \frac{1}{3}\left(\frac{\Lambda - x_i}{\Lambda}\right)^3 - \ldots\right\}. \quad \left[\text{if } \frac{\Lambda - x_i}{\Lambda} < 1\right]$$

$$= log\, \Lambda - \frac{1}{n}\sum\left\{\frac{\Lambda - x_i}{\Lambda} + \frac{1}{2}\left(\frac{\Lambda - x_i}{\Lambda}\right)^2 + \frac{1}{3}\left(\frac{\Lambda - x_i}{\Lambda}\right)^3 + \ldots.\right\}$$

$= log\, \Lambda - a$ positve quantity

or, log G = log Λ- a positive quantity

or, log G ≤ log Λ

or, Λ ≥ G----------(*i*)

Using the relation, $\Lambda \geq G$, we have

$$\frac{x_1 + x_2 + \ldots\ldots + x_n}{n} \geq (x_1.x_2\ldots\ldots x_n)^{1/n}$$

Replacing $x_1, x_2, \ldots, x_n$ by $1/x_1, 1/x_2, \ldots\ldots 1/x_n$ respectively, we get

$$\frac{\frac{1}{x_1} + \frac{1}{x_2} + \ldots\ldots + \frac{1}{x_n}}{n} \geq \left(\frac{1}{x_1}.\frac{1}{x_2}\ldots\ldots\frac{1}{x_n}\right)^{1/n}$$

or, $\frac{1}{H} \geq \frac{1}{G}$

or, G ≥ H.....(ii)

Combining (i) and (ii), we get

$$AM \geq GM \geq HM.$$

The above inequality holds equality if and only if $x_1 = x_2 = \cdots = x_n$.

(email: atique@stat.jnu.ac.bd)

Correlation

Correlation Theory: There are various methods for measuring the relationships existing between variables. The simplest are (1) Correlation Analysis and (2) Regression Analysis.

Correlation: The degree of relationship existing between two or more variables observed from same entity is called correlation.

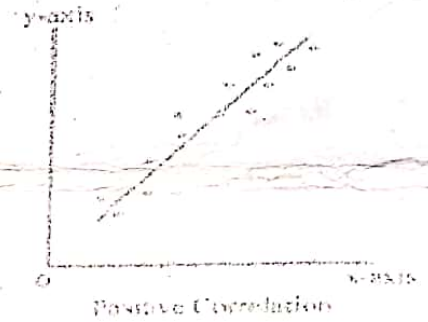
The degree of linear relationship existing between two variables is called simple correlation.

The degree of relationship connecting to three or more variables is called multiple correlation.

Correlation may be linear, when all points (X, Y) on a scatter diagram seem to cluster near a straight line, or non-linear, when all points seem to lie near a curve.

Positive Correlation: Two variables are said to be positively correlated if they tend to change together in the same direction, that is, if they tend to increase or decrease together. Such positive correlation is postulated by economic theory for the quantity of commodity supplied and its price. When the price increases the quantity supplied increases and conversely, when price falls the quantity supplied decreases. The scatter diagram of two variables positively correlated appears in the following figure.

All points in the scatter diagram seem to lie near a line or a curve with positive slope. If all points lie on the line (or curve) the correlation is said to be perfect positive.

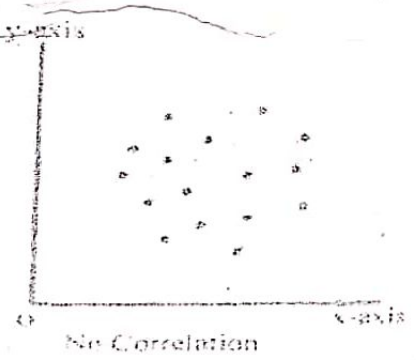


Positive Correlation

Negative Correlation: Two variables are said to be negatively correlated if they tend to change in the opposite direction, that is, when X increases Y decreases, and vice versa. For example, the quantity of commodity demanded and its price are negatively correlated. When price increases, demand for the commodity decreases and when price falls demand increases. The scatter diagram appears in the following figure.

The points cluster around a line (or curve) with negative slope. If all points lie on the line (or curve) the correlation is said to be perfect negative.

No Correlation, or Zero Correlation: Two variables are uncorrelated when they tend to change with no connection to each other. The scatter diagram will appear as in the following figure



The points are dispersed all over the surface of the XY plane. For example, one should expect zero correlation between the height of the inhabitants of a country and the production of steel, or between the weights of students and the color of their hair.

Correlation Coefficient: Correlation coefficient is a quantitative measure of the direction and strength of linear relationship between two numerically measured variables.

For example, if we measure the correlation between X and Y, the population correlation coefficient is presented by ρ_{xy} and its sample estimate by r_{xy}.

Suppose, a random sample of n pairs of values $(X_i, Y_i); i = 1,2,\dots,n$ are given. Therefore, the correlation coefficient between two random variables X and Y, is defined as r_{xy}.

$$r_{xy} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

We can write,

$$\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = \sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2$$

Similarly, $\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2$

$$or, \quad r_{xy} = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right)\left(\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2\right)}}$$

$$or, \quad r_{xy} = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)\left(\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n}\right)}}$$

This formula is also known as Karl Pearson's correlation coefficient of sample.

Properties of r_{xy}:

1) The correlation coefficient is a symmetric measure, that is, $r_{xy} = r_{yx}$.
2) The correlation coefficient will be negative or positive.
3) The correlation coefficient lies between -1 to 1. i.e., $-1 \leq r \leq 1$.
4) The correlation coefficient is a dimensionless quantity, i.e., unit free measurement
5) The correlation coefficient is independent of origin and scale of measurement, i.e., $r_{uv} = r_{xy}$.
6) The correlation coefficient is applicable only for linear relationships.

- **Assumptions of r_{xy} /Pearson's Correlation Coefficient is valid when**

i). Both variables are measured on an interval or ratio scales.
ii). We have sample from bivariate normal population.
iii). The variables involved are quantitative.
iv). Linear relationship between the variables of interest must exist.

Interpretation of Correlation Coefficient:

r =1. This implies that
1. Two variables are perfectly correlated.
2. The relationship between two variables is positive, i.e., if one variable increases, then the other variable also increases.
3. There is a strong linear relationship between the two variables.

r = -1. This implies that
1. Two variables are perfectly correlated.
2. The relationship between two variables is negative, i.e., if one variable increases, then the other variable decreases.
3. There is a strong linear relationship between the two variables.

r = 0.75 This implies that
1. There exists a strong linear relationship between the two variables.
2. The relationship between two variables is positive.

r = -0.75 This implies that
1. There exists a strong linear relationship between the two variables.
2. The relationship between two variables is negative.

r = 0.50 This implies that
1. The linear relationship between two variables is not so high.
2. The relationship between two variables is positive.

r = 0.25 or r < 50 This implies that
1. There exists a less linear relationship between the two variables.
2. The relationship between two variables is positive.

r = 0 This implies that
1. There exists no linear relationship between the two variables.
2. Data sets exhibiting no linearity produce.

Theorem 1: The coefficient of correlation between two variables is independent of origin and scale of measurement.

Proof: Suppose we are given a random sample of n pairs of values $(X_i, Y_i), i = 1,2,\dots,n$.

Then the coefficient of correlation between X and Y is

$$r_{xy} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Let us define, $U_i = \frac{X_i - a}{b} \Rightarrow X_i = a + bU_i$ and $V_i = \frac{Y_i - c}{d} \Rightarrow Y_i = c + dV_i$

Where a, b, c and d are coefficients of which a and c are positive.

Let r_{uv} denotes the coefficient of correlation between U and V, then

$$r_{uv} = \frac{\sum(U_i - \bar{U})(V_i - \bar{V})}{\sqrt{\sum(U_i - \bar{U})^2 \sum(V_i - \bar{V})^2}}$$

We have to prove that $r_{uv} = r_{xy}$.

Now $\bar{X} = a + b\bar{U}$ and $\bar{Y} = c + d\bar{V}$.

Therefore,

$$r_{xy} = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

$$= \frac{\sum(a + bU_i - a - b\bar{U})(c + dV_i - c - d\bar{V})}{\sqrt{\sum(a + bU_i - a - b\bar{U})^2 \sum(c + dV_i - c - d\bar{V})^2}}$$

$$= \frac{bd\sum(U_i - \bar{U})(V_i - \bar{V})}{bd\sqrt{\sum(U_i - \bar{U})^2 \sum(V_i - \bar{V})^2}}$$

Correlation

We know that $S_x^2 = \frac{1}{n}\sum(X_i - \bar{X})^2 \Rightarrow \sum(X_i - \bar{X})^2 = nS_x^2 - - - - - - -(2)$

Similarly, $\sum(Y_i - \bar{Y})^2 = nS_y^2 - - - - - - -(3)$

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

$\Rightarrow \sum(X_i - \bar{X})(Y_i - \bar{Y}) = nS_{xy} - - - - - - -(4)$

Using (2), (3) and (4) in (1), we get,

$$\frac{nS_x^2}{S_x^2} + \frac{nS_y^2}{S_y^2} \pm 2\frac{nS_{xy}}{S_xS_y} \geq 0$$

$$\Rightarrow 2n \pm 2n\frac{S_{xy}}{S_xS_y} \geq 0$$

$$\Rightarrow 1 \pm \frac{S_{xy}}{S_xS_y} \geq 0$$

$$\Rightarrow 1 \pm r_{xy} \geq 0$$

Taking positive sign, $1 + r_{xy} \geq 0, \Rightarrow r_{xy} \geq -1 - - - - - -$ (5)

Taking negative sign, $1 - r_{xy} \geq 0, \Rightarrow r_{xy} \leq 1 - - - - - -$ (6)

From (5) and (6),

$$-1 \leq r_{xy} \leq 1.$$

Example 1: Seven pairs of observations on the variables X and Y are given below:

X:	1.0	2.2	2.8	3.1	3.8	4.5	5.1
Y:	3.1	12.5	12.4	7.6	9.3	14.7	13.4

Find the Correlation Coefficient between X and Y and comments on it.

Solution: The Correlation Coefficient between X and Y is given by

$$r_{xy} = \frac{\sum X_iY_i - n\bar{X}\bar{Y}}{\sqrt{(\sum X_i^2 - n\bar{X}^2)(\sum Y_i^2 - n\bar{Y}^2)}}$$

Now we have to construct the following table

X_i	Y_i	X_iY_i	X_i^2	Y_i^2
1.0	3.1	3.1	1	9.61
2.2	12.5	27.5	4.84	156.25
2.8	12.4	34.72	7.84	153.76
3.1	7.6	23.56	9.61	57.76
3.8	9.3	35.34	14.44	86.49
4.5	14.7	66.15	20.25	216.09
5.1	13.4	68.34	26.01	179.56
$\sum X_i = 22.5$	$\sum Y_i = 73$	$\sum X_iY_i = 258.71$	$\sum X_i^2 = 83.99$	$\sum Y_i^2 = 859.52$

Here, $\bar{X} = 3.21$ and $\bar{Y} = 10.43$. So,

$$r_{xy} = \frac{258.71 - 7 \times 3.21 \times 10.43}{\sqrt{\{85.99 - 7 \times (3.21)^2\}\{959.52 - 7 \times (10.43)^2\}}} = 0.711$$

Example 2: The following data represent the age (x, year) and blood pressure (y, mmHg) of 10 patients:

Age (x):	25	30	35	30	32	40	45	40	36	35
BP (y)	75	80	85	90	95	85	100	90	85	80

Find the relationship between age and blood pressure and comments on your result. Fit a regression line of blood pressure on age and estimate blood pressure of a man aged 60 years. Also calculate the coefficient of determination.

Solution: $r_{xy} = 0.63$

There is a positive relationship between age and blood pressure meaning that if age increases, the BP also increases.

The fitted regression line is

$$y = 58.62 + 0.80x$$

The estimated BP of a man aged x = 60 years is

$$y = 58.62 + 0.80 * 60 = 106.62$$

Regression Analysis

Regression analysis is a statistical technique to study the cause and effect relationship of two or more variables when variables are recorded from each point of a given sample. More precisely, it is a technique to analyse and to estimate the influence of independent variable on dependent variable.

Simple Linear Regression:

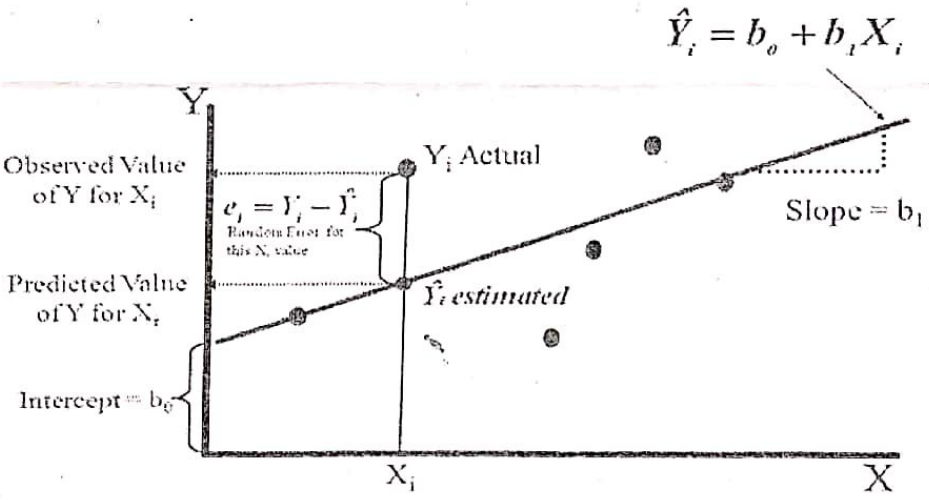
If the dependent variable depends only on <u>one</u> **predictor/explanatory/independent variable** and the relationship is studied by a straight line, the regression is called Simple Linear Regression.

Regression Model: The simple linear regression model is

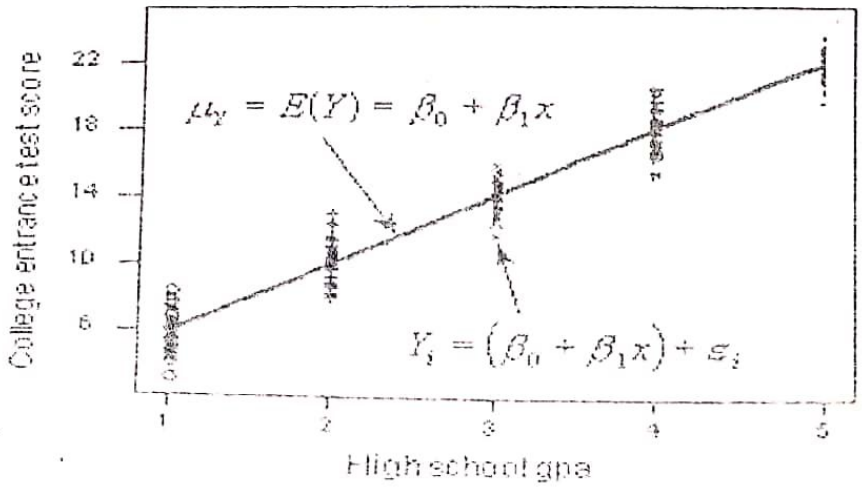
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Where,

- The variable Y is regarded as the **response, outcome**, or **dependent** variable.
- The variable X is regarded as the **predictor, explanatory**, or **independent** variable.
- β_0 is the intercept (parameter) of the regression model measuring **the value of Y in absence of X.**
- β_1 is the regression coefficient of Y on X (β_{yx}) (parameter) which **measures the rate of change of Y for a unit change in the value of X.**
- ε_i is the random error term.



The regression line is depicted by the following figure:



Multiple Linear Regression:

If the **predictor/explanatory/independent variables** simultaneously influence <u>one</u> **dependent** variable, the method of studying the relationship of such predictor/explanatory/independent variables and dependent variables known as multiple regression.

The multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

Assumptions of Regression Analysis:

The assumptions of regression analysis (simple or multiple) are:

(i) The dependent variable $Y's$ are independent of each other.

(ii) The dependent variable Y is assumed to be normally distributed.

(iii) The explanatory variable(s) is/are non-random variable(s)-fixed.

(iv) The explanatory variables are uncorrelated.

(v) The error term ε_i is normally and independently distributed with mean zero and variance σ^2.

Estimation of Parameters in Simple Regression Model/Fitting the Regression Line:

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be n pairs of values observed from a random sample. Assume that the variable X and Y follow the (population) regression model

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \ \ldots\ldots\ldots (1)$$

Assume the fitted regression equation is

$$\hat{y}_i = b_0 + b_1 x_i \ \ldots\ldots\ldots (2)$$

The simplest form of estimating the regression parameters is called the ordinary least squares (**OLS**) estimation. The principle of OLS is minimizing the sum of squared error.

The sum of squares of deviations is given by

$$E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

By solving the normal equations, the value of E will be minimum.

Differentiate E with respect to b_0 and b_1 simultaneously and equating the equations with zero,

$$\frac{\delta E}{\delta b_0} = 0 \text{ and } \frac{\delta E}{\delta b_1} = 0$$

These two equations give

$$\sum_{i=1}^{n} y_i = n b_0 + b_1 \sum_{i=1}^{n} x_i \ \ldots\ldots\ldots (3)$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \ \ldots\ldots (4)$$

The equation (3) and (4) are called normal equations. Solving these two equations, we get

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}}$$

Therefore, the fitted regression equation is given by

$$E(Y|X) = \hat{y}_i = b_0 + b_1 x_i$$

$$\Rightarrow E(Y|X) = \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Coefficient of Determination:

The coefficient of determination measures the proportion or percentage of the total variation in the dependent variable explained by the independent/explanatory variable in the regression model. It is denoted by R^2.

More precisely, the coefficient of determination is a summary measures that tell us how well the sample regression line fits the observed data. It is the measure of the goodness of fit of a regression line.

The coefficient of determination is computed by the following formulae:

$$R^2 = \frac{SSR}{SST}$$

Where, ***SSR*** is the regression sum of squares and ***SST*** is the total sum of squares.

$$SSR = b_1\left\{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right\}$$

$$SST = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}$$

Therefore, the coefficient of determination is

$$R^2 = \frac{b_1\left\{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}\right\}}{\sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n}}$$

Difference between Correlation and Regression Analysis

Comparison	Correlation	Regression
Meaning	Correlation is the degree of linear relationship existing between two or more variables observed from same entity.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and predict response values given the value of the predictor.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the direction and strength of linear relationship between two numerically measured variables.	Regression coefficient indicates the rate of change of dependent variable (Y) for a unit change in the value of independent variable (X).
Limit	$-1 \leq r_{xy} \leq +1$	$-\infty \leq \beta_{yx} \leq +\infty$
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

Example 1: The following data represent the age (x, year) and blood pressure (y, mmHg) of 10 patients:

Age (x):	25	30	35	30	32	40	45	40	36	35
BP (y)	75	80	85	90	95	85	100	90	85	80

a) Find the relationship between age and blood pressure and comments on your result.

b) Fit a regression line of blood pressure on age and estimate blood pressure of a man aged 60 years. Also calculate the coefficient of determination.

Solution: The Pearson's correlation coefficient of age and BP is, $r_{xy} = 0.63$. There is a positive relationship between age and blood pressure meaning that if age increases, the BP also increases.

The fitted regression line is

$$y = 58.62 + 0.80x$$

The estimated BP of a man aged $x = 60$ years is

$$y = 58.62 + 0.80 * 60 = 106.62$$

The coefficient of determination is $R^2 = 0.397$, i.e., 39.7% of the total variation in *BP* explained by the age of the patients in the regression model.

Example 2: The following data represent the values of gestation age (x, days) and birth-weight (y, pound) of some new born babies of 10 mothers:

X	265	250	270	255	260	258	255	265	248	250
Y	6.2	5.8	7.0	5.6	6.0	5.6	6.4	6.8	5.2	6.0

a) Calculate the Pearson's correlation coefficient and comments your result.

b) Fit a regression line of Y on X. Estimate the birth-weight of a baby if his/her gestation age is 280 days. Also calculate the coefficient of determination/goodness of fit of that regression line and comments on it.

Moments: In statistics, moments are certain constant values in a given distribution. The moments help us to determine the nature and form of the underlying distribution.

Moments of a distribution may be calculated from **arithmetic mean** of the distribution or from any **arbitrary chosen value including zero (origin).**

When the moments are computed from **the arithmetic mean** of the distribution, we call them **moments about mean or central moments.**

When they are computed from **an arbitrary value,** we call them **raw moments.** When they are computed from **zero,** they are called **moment about origin.** Moments about origin are also called raw moments.

General Moment:

Let $x_1, x_2, \ldots, x_n$ be n observations of a variable, then the $r-th$ general moment is defined by,

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

where A is any arbitrary value.

Raw Moment/Moments about Origin:

The $r-th$ raw moment is defined by

$$\mu_r' = \frac{1}{n}\sum_{i=1}^{n} x_i^r$$

This is obtained from general moment when $A = 0$.

Central Moment:

The $r-th$ central moment is the mean deviation of $r-th$ order when deviation is measured from $(A = \bar{x})$ mean. Thus, the $r-th$ central moment is given by,

$$\mu_r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^r}{n}.$$

If $x_1, x_2, \ldots, x_n$ occur with frequencies $f_1, f_2, \ldots\ldots, f_n$ respectively then the $r-th$ raw/general moment is

$$\mu_r' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^r}{N}$$

where, $N = \sum_{i=1}^{n} f_i$ and A is any arbitrary value.

And the $r-th$ central moment is

$$\mu_r = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^r}{N}$$

where, $N = \sum_{i=1}^{n} f_i$ and $\bar{x}$ =Arithmetic Mean.

Therefore, the first four raw moments are

$$\mu_1' = \frac{\sum_{i=1}^{n} f_i(x_i - A)}{N}, \quad \mu_2' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^2}{N}$$

$$\mu_3' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^3}{N}, \quad \mu_4' = \frac{\sum_{i=1}^{n} f_i(x_i - A)^4}{N}, \text{ and so on}$$

And the first four central moments are

$$\mu_1 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})}{N} = 0, \qquad \mu_2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N}$$

$$\mu_3 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^3}{N}, \qquad \mu_4 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^4}{N}$$

Example: Compute the first three central moments for the following frequency distribution

X_i:	2	3	4	5	6
f_i:	1	3	7	3	1

Solution: We prepare the following table for computing the moments

x_i	f_i	$f_i x_i$	$(x_i - \bar{x})$	$f_i(x_i - \bar{x})$	$f_i(x_i - \bar{x})^2$	$f_i(x_i - \bar{x})^3$
2	1	2	-2	-2	4	-8
3	3	9	-1	-3	3	-3
4	7	28	0	0	0	0
5	3	15	1	3	3	3
6	1	6	2	2	4	8
Total	15	60	0	0	14	0

$$Here, \quad \bar{x} = \frac{\sum f_i x_i}{N} = \frac{60}{15} = 4$$

$$Thus, \quad \mu_1 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})}{N} = \frac{0}{15} = 0$$

$$\mu_2 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^2}{N} = \frac{14}{15} = 0.933, \quad \mu_3 = \frac{\sum_{i=1}^{n} f_i(x_i - \bar{x})^3}{N} = 0$$

Relationship between raw moments and central moments:

Let X be a random variable assuming values $x_1, x_2, \ldots, x_n$ with mean $\bar{x}$. Let A be any arbitrary value, then

$$\mu_1' = \frac{\sum_{i=1}^{n}(x_i - A)}{n} = \frac{\sum_{i=1}^{n} x_i - nA}{n} = (\bar{x} - A)$$

We know,

$$\mu_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n} = 0$$

$$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n} = \frac{\sum_{i=1}^{n}\{(x_i - A) - (\bar{x} - A)\}^2}{n}$$

$$= \frac{\sum_{i=1}^{n}(x_i - A)^2}{n} - 2\frac{\sum_{i=1}^{n}(x_i - A)}{n}(\bar{x} - A) + (\bar{x} - A)^2$$

$$= \mu_2' - 2\mu_1'.\mu_1' + (\mu_1')^2$$

$$= \mu_2' - (\mu_1')^2.$$

$$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n} = \frac{\sum_{i=1}^{n}\{(x_i - A) - (\bar{x} - A)\}^3}{n}$$

$$= \frac{\sum_{i=1}^{n}(x_i - A)^3}{n} - 3\frac{\sum_{i=1}^{n}(x_i - A)^2}{n}(\bar{x} - A) + 3\frac{\sum_{i=1}^{n}(x_i - A)}{n}(\bar{x} - A)^2 - (\bar{x} - A)^3$$

$$= \mu_3' - 3\mu_2'.\mu_1' + 3\mu_1'(\mu_1')^2 - (\mu_1')^3$$

$$= \mu_3' - 3\mu_2'.\mu_1' + 2(\mu_1')^3.$$

Similarly, $\mu_4 = \mu_4' - 4\mu_3'.\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^4$.

Skewness: Skewness means '**lack of symmetry**' that is departure from symmetry of a distribution. A distribution is said to be skewed if

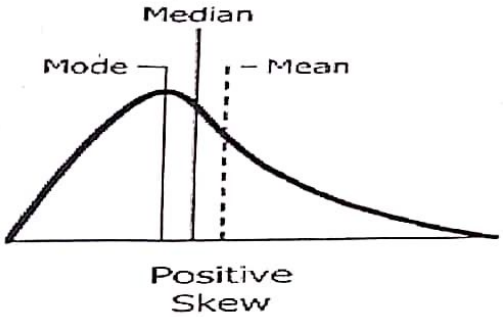
(i) Mean, Median and Mode give different values.

(ii) Q_1 and Q_3 are not equidistant from the median.

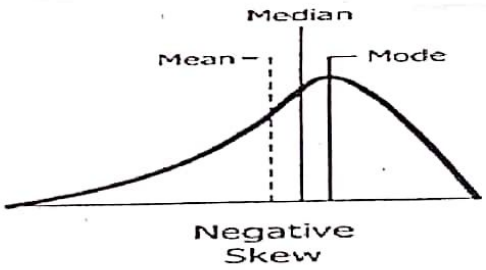
(iii) The curve drawn with the help of the given data is not symmetrical but turned nose to one side than the other.

The following graphs illustrate the skewness of a frequency distribution:

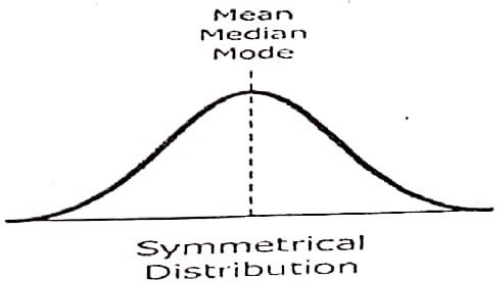
a) Positively Skewed Distribution: Skewness is said to be positive when the tail of the curve of the frequency distribution elongates more on the right. In this distribution, the long tail to the right indicates the presence of extreme values at the positive end of the distribution. This type of a distribution is known as positively skewed distribution. These distributions occur with, for example, family size, female age at marriage, wages of employees etc. For this type of distribution, Mean > Median > Mode.



b) Negatively Skewed Distribution: Skewness is said to be negative when the tail of the curve of the frequency distribution elongates more on the left. In a negatively skewed distribution, the mean is pulled in a negative direction. Daily maximum temperature for a month in winter will result in such a negatively skewed curve. For this type of distribution, Mean < Median < Mode.



c) Symmetrical Distribution: If the curve of the frequency distribution is symmetrical, then skewness is zero. i.e., this type of the distribution is known as normal. For Such a distribution, Mean = Median = Mode.



Why do we study Skewness?

The object of studying skewness is to estimate the direction of which and the extent to which the curve of the frequency distribution is away from the symmetrical distribution.

Measures of Skewness:

- Pearson's Coefficient of skewness is

$$sk = \frac{Mean - Mode}{SD}$$

If Mean > Mode, the skew is positive,

If Mean < Mode, the skew is negative,

If Mean = Mode, the skew is zero, the distribution is symmetrical.

For a moderately skewed distribution, **Mean - Mode = 3(Mean - Median)**

Therefore, Pearson's Coefficient of skewness is

$$sk = \frac{3(Mean - Median)}{SD}$$

If $sk > 0$, the skew is positive

If $sk < 0$, the skew is negative

If $sk = 0$, the skew is zero, the distribution is symmetrical.

Relative measures of skewness in terms of moments:

A relative measure of skewness denoted by β_1, is defined as

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}$$

Instead of β_1, Karl Pearson suggested γ_1 to be used as a measure of skewness

$$\gamma_1 = \sqrt{\beta_1} = \sqrt{\frac{\mu_3^2}{\mu_2^3}} = \frac{\mu_3}{\mu_2^{3/2}}$$

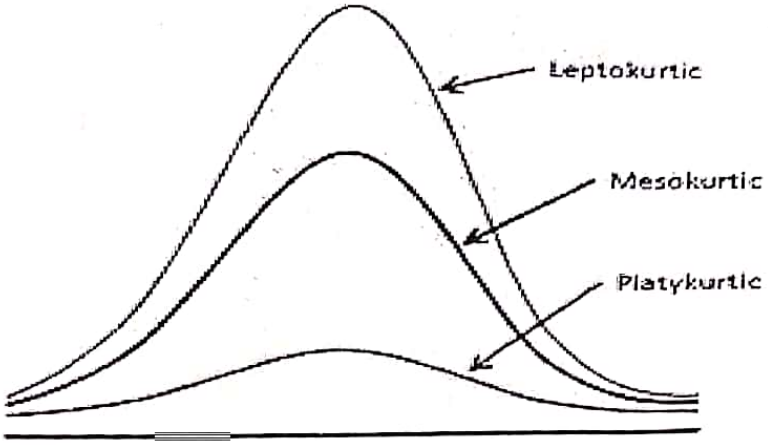
If $\gamma_1 > 0$, the skew is positive

If $\gamma_1 < 0$ *or* $\mu_3 < 0$, the skew is negative

If $\gamma_1 = 0$, the skew is zero, the distribution is symmetrical.

Kurtosis: The degree of peakedness or flatness of a distribution relative to a normal distribution is called kurtosis.

A curve having relatively higher peak than the normal curve, is known as leptokurtic. If the curve is more flat-topped than the normal curve, it is called platykurtic. A normal curve itself is called mesokurtic, which is neither too peaked nor too flat-topped.



Measures of Kurtosis: The most important measure of kurtosis based on 2nd and 4th moments is β_2, defined as $\beta_2 = \frac{\mu_4}{\mu_2^2}$, where, $\mu_2$and $\mu_4$are respectively the second and forth moments about the mean. This measure is a pure number and is always positive.

If $\beta_2 > 3$, the distribution is leptokurtic.

If $\beta_2 < 3$, the distribution is platykurtic.

If $\beta_2 = 3$, the distribution is mesokurtic.

Instead of β_2, Karl Pearson suggested γ_2 to be used as a measure of kurtosis:

$$\gamma_2 = \beta_2 - 3$$

Example: For a distribution, the four central moments were found to be as follows: $\mu_1 = 0$, $\mu_2 = 2.5$, $\mu_3 = 0.7$ and $\mu_4 = 18.75$. Find β_1 and $\beta_2$and hence comment on the nature of the distribution.

Solution:

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(0.7)^2}{(2.5)^3} = 0.031 \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{18.75}{(2.5)^2} = 3$$

Based on the values of β_1 and β_2, we conclude that the distribution is approximately symmetric and mesokurtic.

Theorem: For any set of observations $x_1, x_2, \ldots, x_n$, prove that $(i)\ \beta_2 \geq \beta_1 + 1$ and $(ii)\ \beta_2 \geq 1$.

Proof: (i) We know that,

$$\mu_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}$$

$$\mu_3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{n}, \quad \text{and} \quad \mu_4 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{n}$$

Consider the following expression,

$$\{a(x_i - \bar{x})^2 + b(x_i - \bar{x}) + c\}^2$$

where, a, b, c are arbitrary constants. If a, b *and* c are assumed to be real, then the above is always positive. Thus, performing the squares, summing and dividing by n all through

$$\frac{a^2\sum(x_i - \bar{x})^4}{n} + \frac{b^2\sum(x_i - \bar{x})^2}{n} + c^2 + 2ab\frac{\sum(x_i - \bar{x})^3}{n} + 2ac\frac{\sum(x_i - \bar{x})^2}{n} + 2b\frac{\sum(x_i - \bar{x})}{n} \geq 0$$

$$\Rightarrow a^2\mu_4 + b^2\mu_2 + c^2 + 2ab\mu_3 + 2ac\mu_2 + 0 \geq 0$$

Choosing $a = 1$, $b = -\frac{\mu_3}{\mu_2}$ and $c = -\mu_2$, the above expression becomes,

$$\mu_4 + \frac{\mu_3^2}{\mu_2} + \mu_2^2 - 2\frac{\mu_3^2}{\mu_2} - 2\mu_2^2 \geq 0$$

$$\Rightarrow \mu_4 - \frac{\mu_3^2}{\mu_2} - \mu_2^2 \geq 0$$

Dividing both sides by μ_2^2,

$$\frac{\mu_4}{\mu_2^2} - \frac{\mu_3^2}{\mu_2^3} - 1 \geq 0$$

$$\Rightarrow \beta_2 - \beta_1 - 1 \geq 0$$

$$\Rightarrow \beta_2 \geq \beta_1 + 1.$$

(ii) Let $z_i = (x_i - \bar{x})^2 - - - - - (1)$

summing both sides of (1) and dividing throughout by n

$$\frac{1}{n}\sum z_i = \frac{1}{n}\sum(x_i - \bar{x})^2$$

$$\Rightarrow \bar{z} = \frac{1}{n}\sum(x_i - \bar{x})^2 = \mu_2$$

$$i.e., \bar{z}^2 = \mu_2^2 - - - - - - - - - - - (2)$$

$$Squaring (1)$$

$$z_i^2 = (x_i - \bar{x})^4$$

$$\Rightarrow \frac{1}{n}\sum z_i^2 = \frac{1}{n}\sum(x_i - \bar{x})^4 = \mu_4 - - - - - - - (3)$$

$$Since, \quad \frac{1}{n}\sum(z_i - \bar{z})^2 \geq 0$$

$$\Rightarrow \frac{1}{n}\sum z_i^2 - \bar{z}^2 \geq 0$$

Substituting the values of z_i^2 and $\bar{z}$ from (2) and (3)

$$\mu_4 - \mu_2^2 \geq 0$$

$$\Rightarrow \frac{\mu_4}{\mu_2^2} - 1 \geq 0$$

$$\Rightarrow \beta_2 \geq 1.$$

	Sociology	Physics	Finance	Probability	News
First Moment	Original	velocity $v = \frac{dx}{dt}$	delta $\delta = \frac{\partial c}{\partial S}$	Expected Value $E[x]$	Original
Second Moment	Simulation (Copy)	acceleration $a = \frac{dv}{dt}$	gamma $\gamma = \frac{\partial \delta}{\partial s} = \frac{\partial^2 c}{\partial s^2}$	Variance σ_x^2	Reply RT/MT/QT
Third Moment	Meme (Copy of Copy)	velocity of acceleration $\frac{d^2a}{dt^2}$	delta of gamma $\frac{\partial \gamma}{\partial S}$	Skew γ	Reply to RT/MT/QT Moment *RT/MT/QT=Meme*
Fourth Moment	Simulacra	acceleration of acceleration $\frac{d^3a}{dt^3}$	gamma of gamma $\frac{\partial^2 \gamma}{\partial S^2}$	Kurtosis κ	Original Tweet Referencing Meme