PHT 313 Basic Public Health Statistics I

Origin of statistics

The word statistics seems to have been derived from the Latin word '*status*' or the Italian word '*statista*' or the German word '*statistik*' each of which means a political state. In ancient times, government used to collect the information regarding the population and property or wealth of the country.

Statistics has been defined differently by different authors from time to time. The reasons for a variety of definitions are is, in modern times the field of utility of statistics has widened considerably. In ancient times, Statistics was confined only to the affairs of State but now it embraces almost every sphere of human activity. Hence a number of old definitions which were confined to a very narrow field of enquiry were replaced by new definitions which are much more comprehensive and exhaustive.

Perhaps the best definition seems to be one given by Croxton and Cowden, according to whom Statistics may be defined as '*the science which deals with the collection, analysis and interpretation of numerical data*'.

Public health statistics

There are many fields of applied statistics depending upon the science where it is applied. Biostatistics is one of the field of applied statistics.

Biostatistics is the term used when tools of statistics are applied to the data that is derived from biological sciences such as health research fields including medicine, public health, genetics, laboratory studies and epidemiology. Biostatistics emphasis on problems from medicine, public health and biology.

Areas of statistics

Broadly divided into two major areas: descriptive statistics and inferential statistics depending on the purpose of the study.

Descriptive statistics and inferential statistics are interrelated. Descriptive statistics organizes the collected data in a systematic manner to describe the data set and inferential statistics analyzes the data to produce significant inferences about it.

Descriptive Statistics

Descriptive statistics refers to statistical techniques used to summarize and describe a data set. Descriptive statistics describe the location, the dispersion and the pattern of distribution by using numerals, tabulations or graphs i.e., there are two basic approaches: numerical and graphical.

Inferential Statistics

As the name suggests inferential statistics, it attempts to make an inference about our data. By inferential statistics we can make inferences concerning some unknown aspect of a population from which a sample is drawn. There are two main methods used in inferential statistics: estimation and hypothesis testing.

Application of biostatistics

The field of biostatistics has become indispensable tool as goal for understanding the cause, natural history and treatment of the disease in order to improve human health reducing illness. Some of the applications are presented below:

- To define what is normal of healthy in a population and to find limits of normality in variables such as height, weight, pulse rate etc.
- To find the action of drug – a drug is given to animals or humans to see whether the changes produced are due to the drug or by chance.
- To compare the efficacy of a particular drug.
- To find an association between two attributes such as smoking and cancer etc.
- To test usefulness of sera and vaccine in the field – percentage of attacks or deaths among vaccinated subjects is compared with that among the unvaccinated ones to find whether the difference observed is statistically significant.
- In epidemiological studies – the role of causative factors is statistically tested. Deficiency of iodine as an important cause of goiter in a community is confirmed only after comparing the incidence of goiter causes before and after giving iodised salt.
- Health and vital statistics are essential tools in demography, public health, medicine practice and community services. Recording of vital events in birth and death registers and diseases in hospitals is like book keeping of the community, describing the incidence or prevalence of diseases, defects or deaths in a defined population.
- To determine the success or failure of specific health programs; to promote health legislation; prioritization of health program so that the community's health may be promoted.

Functions of biostatistics

The main functions of biostatistics are given below:

<u>To simplify complexities</u>

Statistics consists of aggregate of numerical facts. Huge facts and figures are difficult to remember. The complex mass of figures can be made simple and understandable with the help of statistical methods.

<u>To present facts in definite form</u>

One of the important functions of statistics is to present the general statements in precise and definite form. The conclusion stated numerically is definite and hence more convincing than the conclusions stated qualitatively.

<u>To estimate for the present and forecast for the future</u>

While preparing suitable policies and plans, it is necessary to have the knowledge of future tendency. Statistical methods provide helpful means in forecasting the future by studying and analyzing the tendencies based on past records.

<u>To facilitate comparison</u>

The science of statistics does not mean only counting but also comparison. Unless the figures are compared with other figures with the same kind, they are meaningless. Statistical methods such

as average, ratio, rates, percentage, coefficients etc offer the best way of comparison between two phenomena which will enable to draw valid conclusion.

To help in formulation of policies
Statistics helps in formulating the policies in different fields. The government policies are also framed on the basis of statistics. In fact, without statistics, suitable policies can not be framed.

To help in formulation and testing the hypothesis
Statistical methods are helpful not only in estimating the present, forecasting the future but also helps in formulating and testing the hypothesis for the development of new theories.

Limitations
Statistics, with its wide applications in almost every sphere of human activity is not without limitations. The followings are some of its important limitations:

Statistics is not suited to the study of qualitative phenomenon.
Qualitative phenomenon can not be expressed numerically are not capable of direct statistical analysis. However statistical techniques may be applied indirectly by first reducing the qualitative expressions to precise quantitative terms.

Statistics does not study individuals
Individual items, taken separately do not constitute statistical data and are meaningless for any statistical enquiry.

Statistical laws are not exact
Unlike physical and natural sciences, statistical laws are only approximation and not exact. Statistical conclusions are not universally true – they are true only on average.

Statistics is liable or be misused
Statistics must be used by experts. The use of statistical tools by in experiences and untrained persons might lead to very fallacious conclusion. As King says "Statistics are like clay of which one can make a god or devil as one pleases."

Collection of data
The following points should be carefully considered before starting the work of data collection:

Objective of the enquiry
It is not possible of collect appropriate data if the objectives of the statistical enquiry are not known. It is therefore, essential to determine objectives beforehand. If the objectives are clear, it becomes quite easy to decide the exact nature of the data to be collected.

Scope of the enquiry
The next thing to be determined is the scope of the enquiry. Scope indicates the geographical region to be covered by the enquiry and the types of data to be collected from that region. For example, if the purpose of the enquiry is to study the health condition of factory workers, the

enquiry may cover all the factories of Nepal or of a zone or only of a city. The type of factory may be small or big is also to be specified.

Source of information

Once the objective and scope of the enquiry are finalized, one has to determine the sources from which the required information are obtained. Generally there are two sources: primary source and secondary source. The data collected for the first time by the investigator for the proposed enquiry are called primary data. Very often the investigator can use the data which have been already collected by others. Then he is used to use the secondary data.

Methods of data collection

When primary data are needed for the enquiry, there are mainly two methods to collect them. They are the census and sample method. In the census method, information about each and every unit of the population are to be obtained. In the sample method, only a part of the population is selected as a representative of the whole. We can draw reasonably accurate inference about the entire population on the basis of information obtained from the sample data. The sampling technique is mostly in use these days as it considerably reduces cost of the enquiry and also the time taken to complete it.

Statistical units

The next important step is to clearly define the statistical unit in which the data are to be measured and collected.

Degree of accuracy

An investigator should decide in advance the degree of accuracy he wants to achieve. He can then collect the data accurate enough to serve his purpose. It is, of course not possible to attain cent percent accuracy in a statistical enquiry. In most of the enquiries, only a reasonable degree of accuracy is good enough to give satisfactory results and more accurate estimates are not of much use. An attempt to get higher degree of accuracy in such case will simply be a waste of time, labor and money. The degree of accuracy needs depends upon the nature of the enquiry and its objective and scope. The available resources and time should also be taken into consideration.

Primary and secondary data

The data required for a statistical enquiry are sometimes collected by the investigator himself or by his agents. These data collected for the first time are in original character and they are called primary data. In most of the statistical enquiries the data already collected are used by others. Such data are called secondary data for him. The distinction between primary data and secondary data is not so rigid. The same data may be primary for some person and secondary when they are used by others.

Methods of collecting primary data

1. Experiments

Experiments are performed in the various laboratories for investigations. The data collected with specific objective by one or more workers are compiled and analyzed. These results are made use of in preparation of dissertation, theses and scientific papers of publication in scientific journals and books.

2. Survey

Surveys are carried out for epidemiological studies in the field by trained teams to find the incidence or prevalence of health or disease situations in a community etc. The following are the various methods used to conduct survey to generate data:

Direct personal interview

In this method the investigators himself goes to the spot and makes direct personal contacts with the persons who are directly concerned with the problem under study. The data obtained by this method are also accurate because the interviewer can clear up any doubt of the informants about the meaning or purpose of the questions.

Indirect oral interview

In this method, the investigator does not make any contact with the persons who are directly concerned with the problem under study. The required data are obtained by interviewing third persons who are in touch with the field of enquiry and can give the desired information.

Information from correspondents

In this method data are not collected by the investigators or enumerators. Some agents known as correspondents are employed in various places. They collect information in their own ways and send them to the central office where the data are processed and analyzed. This method is generally used by newspaper to get information.

Mailed questionnaire method (Questionnaire through post)

In this method a questionnaire relating to the enquiry is prepared. It contains a list of questions and some space for answers to be filled by the respondents. Copies of the questionnaire are sent by post to various persons at different places. They are requested to answer all the questions and return it by post within a specified time.

Schedules to be filled in by enumerators

Just like a questionnaire a schedule also contains a set of questions. While the questionnaire is sent by post and filled in by the respondent, the schedule is sent through the enumerator who fills it up. The enumerator is specially trained and appointed for the purpose. He goes to the informants with the schedule and meets them personally and records the answers given by filling up the schedules himself.

Self administered

A tremendous amount of recent data is readily available through the Internet. These include already tabulated observations and reports, as well as access to the raw data. Some of this data cannot be accessed through the Internet because of confidentiality requirements, but even in this case it is sometimes possible to get permission to analyze the data at a secure site under the supervision of employees of the agency.

Sources of secondary data

It is always not possible to collect the data from primary source due to time, money and labor. So in many situations we have to take already collected data from any sources for our purpose. The

data which is compiled for our purpose from any source is known as secondary data. There is a lot of published and unpublished information from which the further studies can be made.
The sources of secondary data are:

- published source
- unpublished source.

Published source
The main sources of published data are the records which are obtained from:

- Official publications of ministries, (health related data can be obtained from Ministry of Health and Population), Central Bureau of Statistics, departments of governments.
- Official publications of international agencies like UN, FAO, WHO etc.
- Reports and publication of reliable organization, journals, periodicals, newspaper, research works (research papers),
- Publication of research institutes,
- Electronic resources.

Unpublished source

- Medical records
- Administrative data: referrals, sites of care.
- Vital records: records of birth and death etc.;
- Records of disease registries;
- Data of specific health programs;
- Thesis and dissertation of any academic research.

Measurement and scale
In powerpoint slides

Common Measures of Health Status
In public health, vital statistics, epidemiology, and other aspects of the health sciences and health care, events of interest include birth, deaths, outbreaks of disease, spells of sickness, hospitalizations, immunizations against infectious diseases, and many other events and phenomena has to be taken in account.
In epidemiology, many nominal variables have only two possible categories: alive or dead; case or control; exposed or unexposed etc. Such variables are called dichotomous variables.
The epidemiologist usually expresses disease magnitude as a count, rate, ratio and proportion. A clear understanding of the term is required for proper interpretation of epidemiological data.

Count
Count is the simplest and most basic measure. It is a absolute number of persons who have disease or characteristic of interest. It is useful for health planners and administrators: for allocation of resources (e.g. quantity of ORS needed by diarrheal cases). To measure incidence or prevalence of the condition, we usually want to count people. To measure the public health burden, we may want to count events. For example, count of number cases of a disease, is used for surveillance of infectious disease for early detection of outbreaks.
To compare frequencies of phenomena at different times and places, among different classes of persons we use a term rate which is more than a number.

Rate

Much of public health assessment involves describing the health status of a defined community by looking at changes in the community over time or by comparing health events in that community to events occurring in other communities or the state as a whole. In making these comparisons, we need to account for the fact that the number of health events depends in part on the number of people in the community. To account for growth in a community or to compare communities of different sizes, we usually develop rates to provide the number of events per population unit.

According to the dictionary, a rate is, "a quantity, amount, or degree of something (numerator), measured per unit of something else (denominator)." In public health, the numerator is the number of people among whom an event occurred during a certain period of time, and the denominator is the total number of people in the population at risk for the same period of time. For example, in looking at female breast cancer, we cannot include men in the population at risk, because men with breast cancer would not be included in the numerator.

A rate has four components:

1. A specified time period,
2. Numerator (the number of events),
3. Denominator (the population at risk of experiencing the event),
4. A constant. The result of the fraction is usually multiplied by some factor of 10, so that the rate may be expressed as a whole number

For example,

$$\text{Rate} = \frac{\text{Numerator}}{\text{Denominator}} \times \text{Constant multiplier}$$

$$\text{Crude death rate} = \frac{\text{number of deaths during a year}}{\text{mid year population}} \times 1000$$

$$\text{Case fatality rate} = \frac{\text{total number of deaths due to a particular disease}}{\text{total number of cases due to the same disease}} \times 100$$

Ratio

Another measure of disease frequency is a ratio. It expresses a relation in size between two random quantities. The numerator is not a component of the denominator. The numerator and denominator may involve an interval of time or may be instantaneous in time. Broadly, ratio is the result of dividing one quantity by another. It is expressed in the form of x:y or $\frac{x}{y}$.

For example, $\text{Fetal death ratio} = \frac{\text{fetal death}}{\text{live birth}}$

Fetal deaths are not included among live births, by definition.

Some examples of ratio are:

- the ratio of white blood cells relative to red blood cells is 1:600 means for each white cell there are 600 red cells.
- doctor – population ratio, Number of doctor per beds (1 doctor for 50 beds)
- child – woman ratio.
- Number of participants per facilitator

- Sex ratio: Females / Males
- Number of children with scabies / number of children with malnutrition

Proportion

When two ratios are equal, the mathematical statement of that equality is called a proportion. To find out if two ratios form a proportion, one can evaluate the cross product. If $\frac{a}{b}$ and $\frac{c}{d}$ are ratios, then the two ratios form a proportion if ad = bc.

A proportion is a ratio which indicates the relation in a magnitude of a part of the whole. The numerator is always included in the denominator i.e. numerator is a part of denominator. A proportion is usually expressed as a percentage. Most fractions in epidemiology are proportions. All rates are ratios; some rates are proportions (the numerator is a portion of the denominator). For example, the "case fatality rate" is the number of fatal cases as a proportion of all cases of a condition, such as hospitalized patients with acute heart attacks.

Variable

The quantitative characteristics which can take different values under study is termed as variable. Age, weight, height, pulse rate, blood pressure etc. are some examples of variable.

The value which the variable takes is known as variate values. There are two types of variable: discrete and continuous.

The variable which can take countable or finite or fixed number of values is known as discrete variable. For example: no. of drug in a strip, no. of students in a class, no. of patients arrived in a hospital, no. of patients visited in a laboratory, number of tests performed in a laboratory etc.

The variable which can assume to take all the possible values in the given range in known as continuous variable. For example: height, weight, systolic blood pressure etc.

Frequency

The number of times which a variate value occurs is known as frequency. the variate value 5 is repeated 3 times means the frequency of the variate value 5 is 3.

Frequency distribution

The arrangement of variate values along with frequency in the tabular form is known as frequency. Frequency distribution is simply a table in which the collected and classified data are presented into different classes and the number of cases which fall in each class are recorded. Frequency distribution is of two kinds: univariate and bivariate.

Again univariate frequency distribution may be of following three types: Individual series, discrete series, continuous series.

Individual series

In this series, items are listed singly after observations, as distinguished from listing them in groups. Example**:**

Age in years of 5 patients in a certain hospital is recorded as: 69, 60, 57, 68, 30

Day	Sun	Mon	Tue	Wed	Thurs	Fri	Sat
Temp.(^{o}c)	32	31	26	25	35	28	24

Discrete series

The series formed by discrete variable is known as discrete series. This series is made if the data are in small size. In a discrete series, the data are presented in a way that exact measurements of units are clearly indicated. There is a definite difference between the variate values.

Example: Days of confinement after delivery is given below:

Days of confinement	6	7	8	9	10
No. of patients	5	4	4	3	2

Continuous series

The series formed from a continuous variable is termed as continuous series. In continuous series, the data or variate values are always between two numbers known as interval (class). So in continuous series the class interval are given along with the frequency.

Example: the age distribution of a certain disease reported during a year in a particular state is presented below:

Age	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
No. of cases	5	10	120	22	13	5

Construction of discrete frequency distribution

The main objective of preparing a discrete frequency distribution is to change the given raw data into condensed and simpler form. The following are the steps to be used in preparing the discrete frequency distribution.

1. In the first column, arrange the variate value in ascending or descending order of magnitudes. Such an arrangement is known as array.
2. In a second column put a vertical bar (|) known as tally bar to count the variate values.
3. Count the number of tally bars corresponding to each of the variable and place it in the third column headed frequency.

Methods of forming grouped frequency distribution

Exclusive method

The formation of class interval by this method is that the upper limit of lower class is same as that of lower limit of next class so as to make continuous without any gap. In this method the upper limit of class is excluded while entering the data. This method is mostly useful in case of continuous variable.

Example: The following distribution of 727 deaths from Cholera (for various ages) classified according to age:

Age	0-5	5-10	10-15	15-20	20-25	25-30	30-45	45-70
No. of deaths	322	213	70	27	26	17	33	19

In the age group 0 – 5, the observations which lie between 0 to 4.999 will lie in this class and when the observation is exactly 5 it will lie in the class 5 – 10 and so on. Since the upper limit of each class is excluded from each class this method is known as exclusive method.

Inclusive method

The formation of the class interval by this method is that both the lower and upper limit of the class is included in the particular class. This method is mostly used in case of discrete variable.

Example: The following distribution of 727 deaths from Cholera (for various ages) classified according to age:

Age	0-4	5-9	10-14	15-19	20-24	25-29	30-39	40-49
No. of deaths	322	213	70	27	26	17	33	19

In this method, there is a gap between the upper limit of the lower class and lower limit of the succeeding class.

Conversion of inclusive to exclusive class

To change the inclusive class to exclusive, we adjust and get the real limits of the classes by using correction factor. The steps are as follows:

1. Find the correction factor.

$$\text{Correction factor (c)} = \frac{\text{lower limit of second class - upper limit of first class}}{2}$$

2. Subtract the correction factor from all the lower limit of the class and add this correction factor all the upper limit of the class to get real class limits.
   Real lower limit = lower limit – c
   Real upper limit = upper limit + c

Conversion of mid values into classes

The steps are as follows:

1. Find the correction factor.

$$\text{Correction factor (c)} = \frac{\text{Difference of two consecutive mid values}}{2}$$

2. Subtract the correction factor from the mid value to get the lower limit of the class and add this correction factor to the mid value to get the upper limit of the class.
   Real lower limit = mid value – c
   Real upper limit = mid value + c

Construction of grouped frequency distribution

The following are the steps to be considered in the construction of discrete frequency distribution.

1. The class should be clearly defined. (whether it is inclusive or exclusive)
2. The number of classes should lie between 5 to 15. If it is less than 5, the necessary details may be lost and if it greater than 15, the further processing of data would be difficult. Prof. Struges' suggested a relation to find the number of class.
   Number of class (k) = $1 + 3.322 \log_{10} n$
   Where n is the number of observations.
3. The magnitude of class interval can be decided by the relation

$$\text{Magnitude (h)} = \frac{\text{Range}}{k} = \frac{L - S}{k}$$

   Where L = largest observation, S = smallest observation
4. After making class, class frequencies should be obtained by using tally bars as in discrete frequency distribution.

Cumulative frequency distribution

The frequency distribution in which the frequencies are cumulated is said to be cumulative frequency distribution. In such distributions, class frequencies are not the individual class

frequencies but cumulated class frequencies. There are two types of cumulative frequency distribution.

Less than cumulative frequency distribution

After arranging the classes (variate values) in ascending order of magnitudes, the frequencies are cumulated from the top. Frequencies thus cumulated are said to be less than cumulative frequency and instead of writing both the limits of the class, we write less than and upper limit of the class.

Example: We observe the following simple frequency distribution.

Birth weight (kg)	No. of children
1.0 – 1.4	12
1.4 – 1.8	12
1.8 – 2.2	7
2.2 – 2.6	10
2.6 – 3.8	9

The change of this distribution to less than cumulative frequency distribution as follows:

Birth weight (kg)	No. of children	Cumulative frequency
Less than 1.4	12	12
Less than 1.8	12	12 + 12 = 24
Less than 2.2	7	24 + 7 = 31
Less than 2.6	10	31 + 10 = 41
Less than 3.8	9	41 + 9 = 50

More than cumulative frequency distribution

After arranging the classes (variate values) in ascending order of magnitudes, the frequencies are cumulated from the bottom. Frequencies thus cumulated are said to be more than cumulative frequency and instead of writing both the limits of the class, we write more than and lower limit of the class. The more than cumulative frequency distribution for above example is as follows:

Birth weight (kg)	No. of children	Cumulative frequency
More than 1.0	12	38 + 12 = 50
More than 1.4	12	26 + 12 = 38
More than 1.8	7	19 + 7 = 26
More than 2.2	10	9 + 10 = 19
More than 2.6	9	9

Relative frequency

$$\text{Relative frequency} = \frac{\text{frequency of a particular class (or variate values)}}{\text{total frequency}}$$

The relative frequency for the above example is,

Birth weight (kg)	No. of children	Relative frequency
1.0 – 1.4	12	$\frac{12}{50} = 0.24$
1.4 – 1.8	12	$\frac{12}{50} = 0.24$

1.8 – 2.2	7	$\frac{7}{50} = 0.14$
2.2 – 2.6	10	$\frac{10}{50} = 0.2$
2.6 – 3.8	9	$\frac{9}{50} = 0.18$
Total	50	

Note: Total of relative frequency = 1

Exercise:

1. The length of stay (days) in hospital for 20 patients admitted for a certain condition were: 54, 55, 58, 59, 64, 55, 73, 62, 73, 64, 55, 59, 64, 58, 64, 67, 73, 64, 62, 65. Construct a discrete frequency distribution.
2. The age in years of 30 patients in a certain hospital is recorded as,
   69 60 57 68 60 65 60 30 44 59 64 65
   40 63 60 45 57 65 56 50 70 73 73 49
   60 52 62 47 51 40 67
   Construct a discrete frequency distribution.
3. In an enquiry of 30 families, the number of children recorded is as follows:
   2 4 2 1 2 3 0 5 2 3 0 4
   2 2 1 2 3 2 4 2 3 1 4 4
   2 0 5 3 3 1
   Represent the above data into discrete frequency distribution. Also construct a relative frequency distribution.

4. The height of forty boys of a class is given below in centimeter. Construct a discrete frequency table.
   130 140 133 135 140 135 132 142 133 132 135 130
   140 133 135 132 130 133 130 133 135 140 132 140
   132 135 133 132 142 130 130 142 140 133 142 133
   133 140 133 142
5. The birth weight (in kilograms) of 30 children is recorded as:2.0, 2.1, 2.3, 3.0, 3.1, 2.7, 2.8, 3.5, 3.1, 3.7, 4.0, 2.3,3.5, 4.2, 3.7, 3.2, 2.7, 2.5, 2.7, 3.8, 3.1, 3.0, 2.6, 2.8, 2.9, 3.5, 4.1, 3.9, 2.8 and 2.2. Construct a grouped frequency distribution.
6. Following still birth rates per 1000 total births were repeated by 30 towns in 2005
   27, 28, 40, 32, 30, 36, 25, 29, 30, 29, 26, 30, 20, 35, 32, 36, 37, 29, 29, 42, 32, 27, 35, 36, 29, 33, 27, 41, 49, 34.
   Prepare a frequency distribution using class interval such as 20-24, 25-29 and so on.
7. The regulation board of health in a particular state specify that the fluoride level must not exceed 1.5 ppm (parts per million). The 25 measurements below represent the fluoride level for a sample of 25 days. Although fluoride levels are measured more than once per day, these data represent the early readings for the 25 days sampled.
   .75 .86 .84 .85 .97 .94 .89 .84 .83 .89 .88
   .78 .77 .76 .82 .71 .92 1.05 .94 .83 .81 .85
   .97 .93 .79
   Construct the frequency and relative frequency distributions for the data.

8. The details of daily protein intake (gm) of 60 adult males are given below:

65	68	12	22	63	43	32	43	42	25	49	27
27	74	38	49	28	42	51	30	36	36	27	23
19	31	42	28	32	28	50	46	30	38	31	79
27	28	21	43	49	16	25	22	23	45	24	12
69	12	24	25	57	47	44	51	23	16	55	65

   Prepare a frequency distribution using class intervals such as 10 – 20, 20 – 30 and so on.
9. The arm spans of 30 adult male employees of a certain firm were measured. They were rounded to the nearest centimeter and recorded.

177	173	183	178	186	168	172	178	171	186	172
179	183	179	169	180	170	186	180	181	179	184
174	181	187	169	177	184	170	174			

   Draw a frequency table with 10 classes by inclusive method taking the first class as 168 – 171. Also show the relative frequencies and cumulative frequencies.
10. Following is the data on the waiting time in minutes for a patient suffering from a common disease for treatment.
    30, 10, 16, 20, 8, 16, 12, 21, 33, 27, 28, 17, 15, 22, 29, 21, 15, 9, 13, 40, 45, 16, 18, 25, 33, 17, 31, 37, 10, 28, 9, 12, 15, 35, 27, 13, 18, 16, 19
    Prepare a statistical table with the following characteristics.
    a) Exclusive frequency distribution.
    b) Inclusive frequency distribution
    c) Less than cumulative frequency distribution
    d) More than cumulative frequency distribution

11. The following table represents the age distribution of 50 breast cancer patients admitted to central cancer hospital.
    73, 76, 66, 65, 58, 60, 52, 51, 39, 41, 83, 73, 65, 57, 50, 52, 52, 53, 79, 81, 82, 41, 60, 77, 67, 77, 68, 44, 60, 45, 54, 62, 49, 56, 55, 47, 44, 46, 54, 63, 63,70, 71, 61, 63, 72, 79, 72 , 79, 73
    From these data construct:
    a) a frequency distribution with five class interval,
    b) a relative frequency distribution,
    c) a cumulative frequency distribution,
    Also find percentage of patients is below 65 years of age.

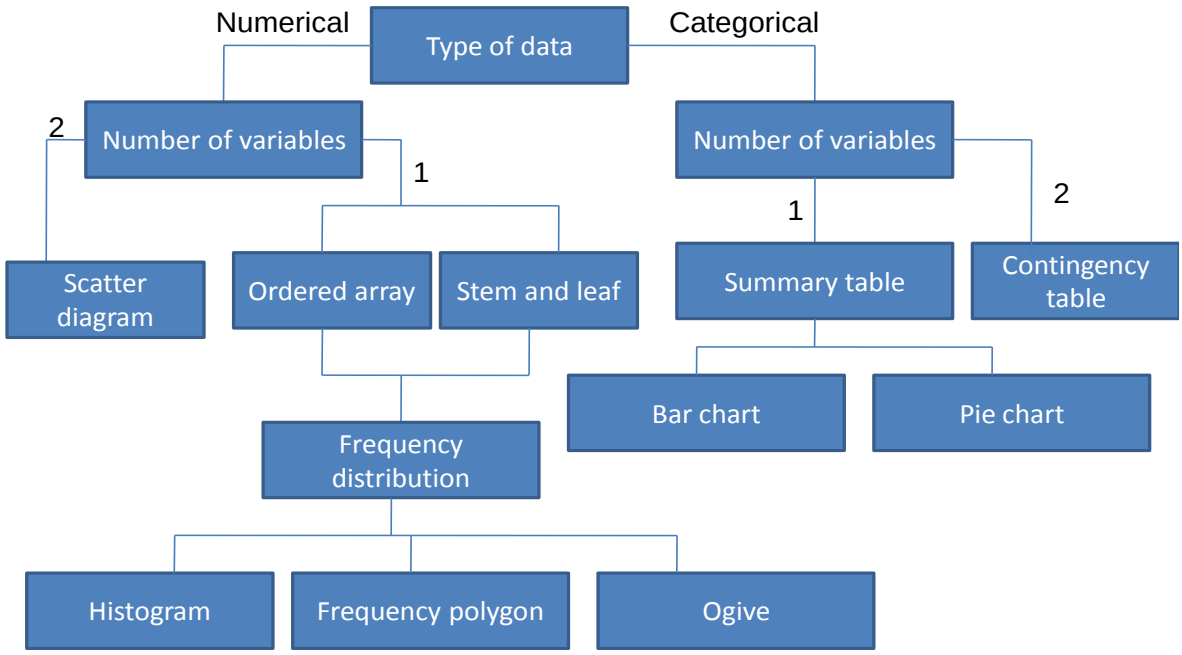
12. The following are the ages of 30 patients seen in the medicine department of hospital on a Friday morning.
    60, 56, 64, 45, 45, 22, 12, 22, 10, 21,53, 44, 54, 32, 36, 35, 36, 23, 38, 55, 35, 43, 45, 46, 57
    From these data construct:
    a) a frequency distribution with five class interval.
    b) a relative frequency distribution
    c) a cumulative frequency distribution.
    Also find percentage of patients is below 40 years of age.

Presentation of data
There are three broad ways of presentation of data. They are: textual presentation, tabular presentation, diagrammatic and graphic presentation.

Textual presentation
In textual presentation the data is presented along with the text. The data presented in textual form is not comprehensive and its significance could not be understood easily. The readers have to go through the whole text again and again to grasp the comprehensive meaning of the text and data. For example, there were 23 cases of measles, 16 cases of diphtheria, 72 cases of kala-azar, gastroenteritis cases were 258 and 22 cases of tuberculosis at a clinic over the past six months among 391 patients visited.

Tabular presentation
Data are classified according to their similarities in different groups in tabular form. *See powerpoint slides.*

Diagrams and graphs
The data obtained are although tabulated neatly, makes confusion and uninteresting in the increase of size of information. One of the important methods of presenting such statistical data is the graphical method. Graphs and diagrams of the data are in the forms of lines, bars, rectangles, circles etc.
The following chart gives a clear idea for the construction of diagrams and graphs.



Types of diagrams

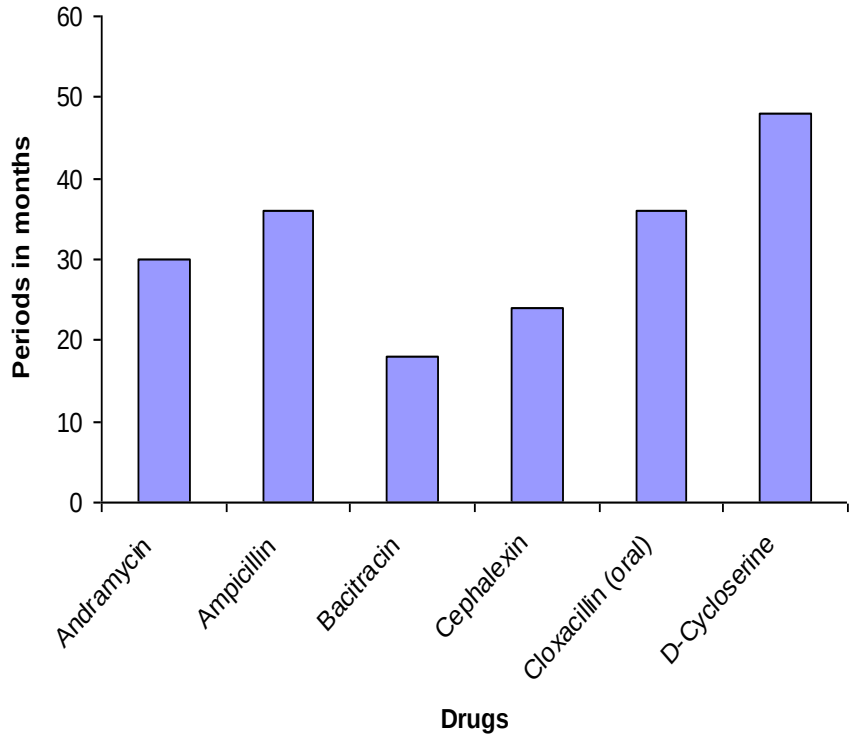
Bar diagram

(i) Simple bar diagram
(ii) Multiple bar diagram
(iii) Sub divided bar diagram

Simple bar diagram

In simple bar diagram, the heights of the bars changes with the change in data. It consists of bars or rectangles. They are made of equal width on a common base. Each bar is kept apart from other. Spacing between any two bars should be nearly equal to half of the width of the bar. The bars may be vertical or horizontal. Generally vertical bars are taken in practice. They are more attractive and are easier to compare. So this is the one of the simplest and most commonly used diagram**.**

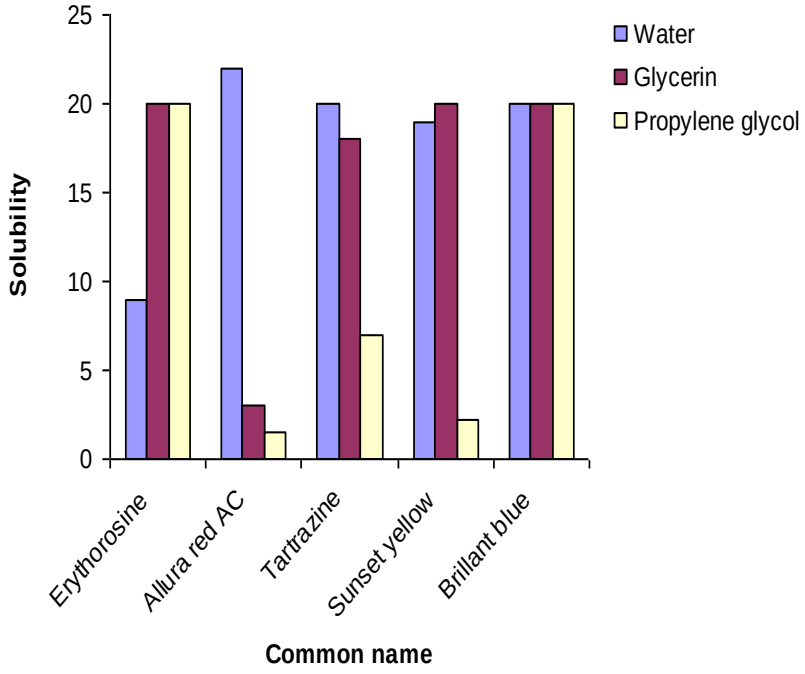
For example: Simple bar diagram showing life periods of drugs



(i) Multiple bar diagram

If two or more sets of interrelated phenomenon or variables are to be presented graphically, multiple bar diagram is used. The technique of drawing multiple bar diagram is basically same as that of simple bar diagram. In this case a set of adjacent bars is drawn. Proper and equal spacing is given between different sets of the bars. To distinguish between different bars in set, different colours, shades may be used and index to this effect may be given.

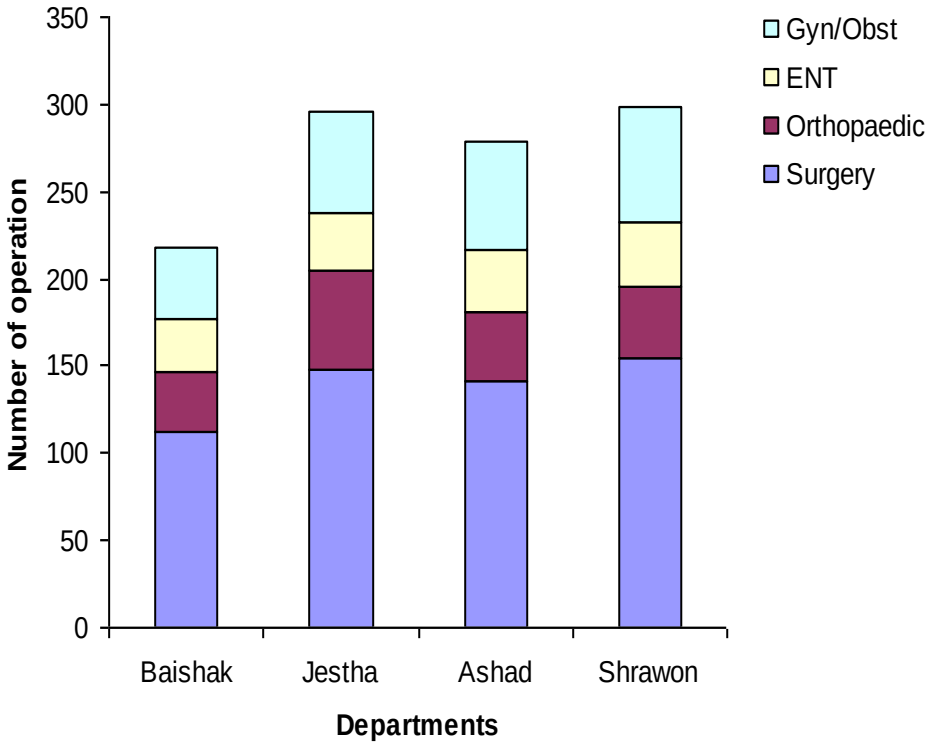
For example: Multiple bar diagram showing the solubility of pharmaceutical colorants



Subdivided bar diagram

Subdivided bar diagrams are useful not only for presenting several items of a variable or category graphically but also enables us to make comparative study of different parts or components among themselves and also to study the relationship between each component and a whole. Subdivided bar diagram is to be used if the total magnitude of the given variable is to be divided into various parts or components. First of all a bar representing the total is drawn. Then it is divided into various segments, each segment representing a given component of the total. Different shades or colours are used to distinguish the various components and an index is given along with the diagram to explain these differences.

For example: Department wise total operations done in the first four months



Pie chart (Angular diagram)

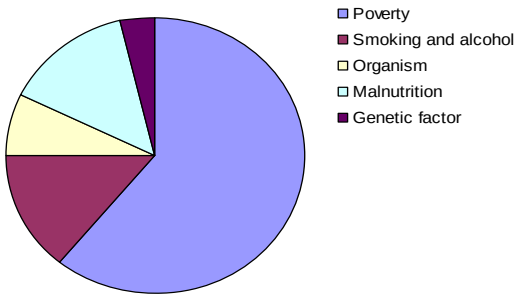
Pie diagram is another popular and widely used method for diagrammatical presentation. It is a two dimensional diagram. It is also known as pie chart or circular diagram or angular diagram. In this method all the given values are converted in terms of angles so that the sum of all these angles equal to 360. The given values are converted into the angles using the following relation.

$$\text{Angle express in degree} = \frac{\text{value of that item}}{\text{total value}} \times 360^{o}$$

To draw two pie charts at the same time of two different data for comparison, we will take a radius of one pie chart and the radius of other is calculated by the formula,

$$\frac{\text{Square root of total of first series}}{\text{Square root of total of second series}} = \frac{\text{radius of first series}}{\text{radius of second series}}$$

For example: Causes of tuberculosis explained by the headings is given below.



Graphs of frequency distribution

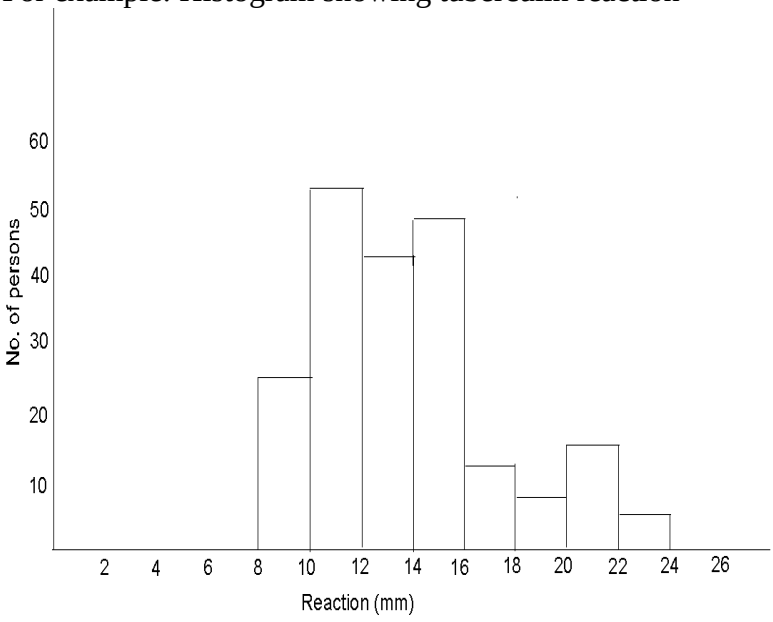
Graphs plotted for frequency distributions are known as frequency graphs. Graphs which are commonly used for frequency distribution are:

i. Histogram
ii. Frequency polygon
iii. Frequency curve
iv. Cumulative frequency curve or 'O give'.

Histogram

A histogram is a series of rectangles, each proportional in width to the range of values within a class and proportional in height to the number of items falling in the class. Histogram is used to describe numerical data which have been grouped into frequency. Variable under consideration is kept along X axis and frequency per class interval is kept in vertical Y axis. If the class intervals on the frequency distribution under consideration are of equal width then the vertical bars in the histogram are also of equal width. If the width of all class intervals is not of equal width, then all class intervals are to be made with equal width class intervals having convenient and suitable class size and the corresponding frequencies are also to be adjusted. The basic assumption that all the frequencies are uniformly distributed throughout the class interval is used while adjusting the frequencies.

For example: Histogram showing tuberculin reaction

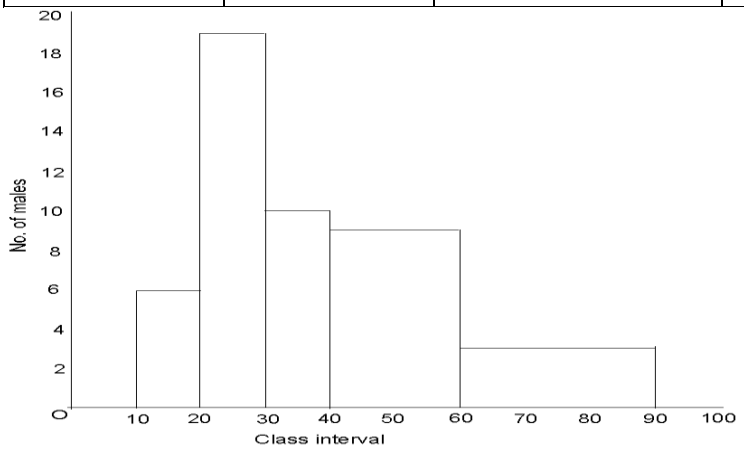


Example: Construct a histogram from the frequency distribution of daily protein intake (gm) of 62 adult males.

Class interval(gm)	10 – 20	20 – 30	30 – 40	40 – 60	60 – 90
No. of males	6	19	10	18	9

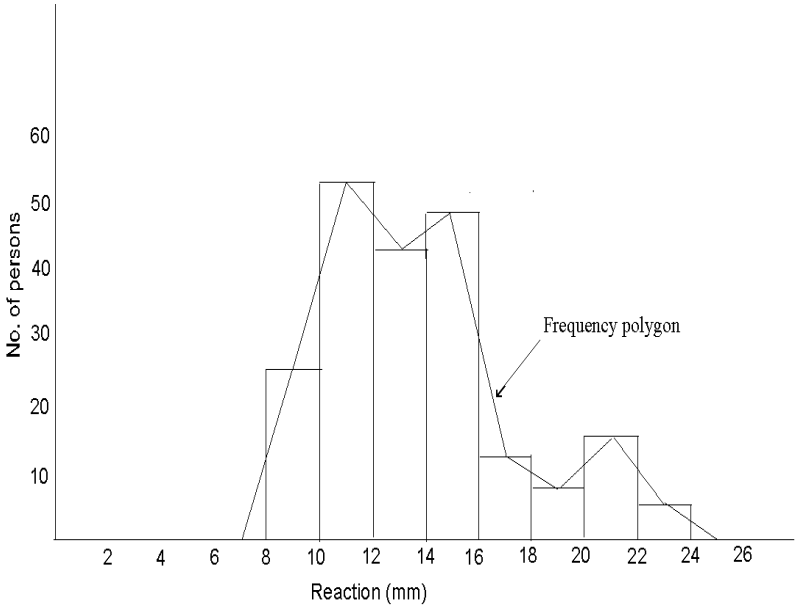
Solution: The given class intervals are of unequal width. Now the frequencies are adjusted accordingly as shown below.

Class interval	No. of males	Magnitude of class	Adjusted frequency
10 – 20	6	10	6
20 – 30	19	10	19
30 – 40	10	10	10
40 – 60	18	20	18/2 = 9
60 – 90	9	30	9/3 = 3

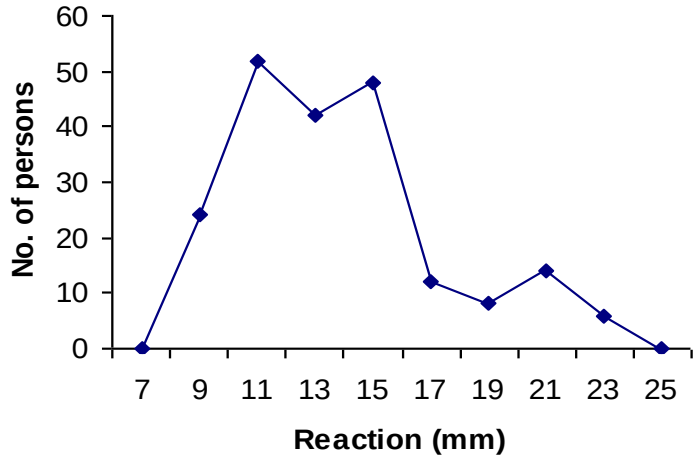


Frequency polygon

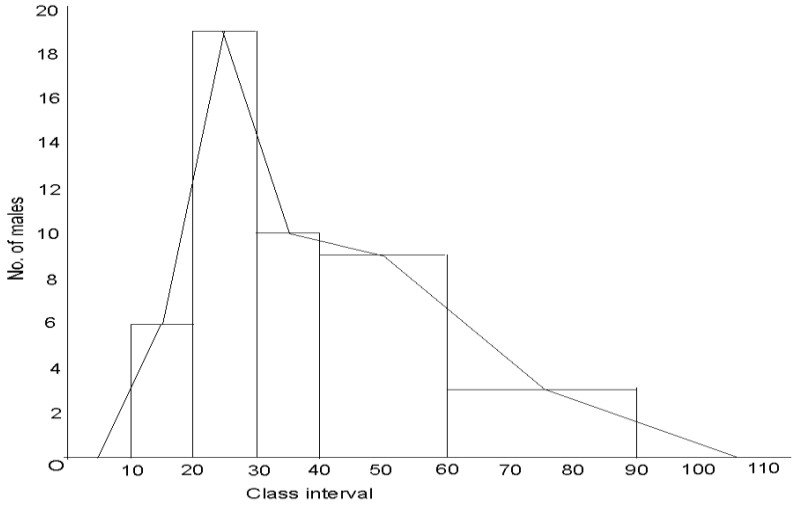
To construct a frequency polygon, a histogram is drawn from the given frequency distribution. The mid points at the top of the rectangles are marked and joined with a straight line. From the mid point of top of the first and last rectangles the straight lines are produced to meet them on x axis such that a closed polygon is formed. The polygon formed this method has the area equal to that of histogram.



Frequency polygon can also be drawn without use of histogram. The mid point of the class interval along with the corresponding frequency is plotted on the graph and the end points are joined on the X-axis as before to get frequency polygon. Frequency polygon without histogram for the above example is shown below.



Graph showing histogram and frequency polygon for above example of unequal class magnitude.



Frequency curve

Frequency curve is obtained by joining the vertices of a frequency polygon or the vertices of mid points of histograms by smooth freehand curve. The main purpose of making frequency polygon is to eliminate the random or erratic fluctuations present in the data. Frequency curve is applicable to find the rate of increase or decrease in frequency. It also helps to estimate the frequencies for the given value of the variable.

Cumulative frequency curve

A graphical presentation of a cumulative frequency distribution is called Ogive or cumulative frequency curve.

Less than Ogive

To use this method, first of all we have to prepare less than cumulative frequency The data are then plotted taking the upper limit of the class interval on the x-axis and cumulative frequency on y-axis. The point thus obtained is joined freely will give the less than Ogive.

More than Ogive

To use this method, first of all we have to prepare more than cumulative frequency. The data are then plotted taking the lower limit of the class interval on the x-axis and cumulative frequency on y-axis. The point thus obtained is joined freely will give the more than Ogive.

Example: Diastolic blood pressure in persons aged 25 to 34 is given below.

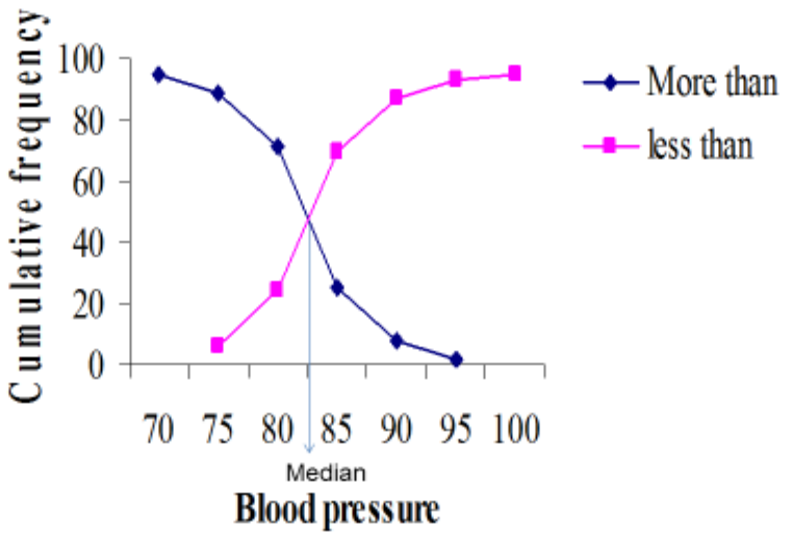
Blood pressure	70 – 75	75 – 80	80 – 85	85 – 90	90 – 95	95 – 100
No. of males	6	18	46	17	6	2

Draw less than and more than ogive.

Solution:

Blood pressure	Less than c.f	Blood pressure	More than c,f
Less than 75	6	More than 70	95
Less than 80	24	More than 75	89
Less than 85	70	More than 80	71
Less than 90	87	More than 85	25
Less than 95	93	More than 90	8
Less than 100	95	More than 95	2

The less than and more than ogive is shown below.



89

Stem and leaf display

In order to organize the large number of data values, basically ordered array or frequency distribution or stem and leaf display is used. Stem and leaf display or stem and leaf plot is a method to organize the statistical data. Stem and leaf display is helpful to see how the data are distributed and where the concentration of the data occur. Thus stem and leaf display helps to have an idea about the shape of the distribution of the given data values.

In stem and leaf plot, each data value is split into a 'stem' and 'leaf'. A vertical line is drawn to separate stem and leaf. The given data is separated in such a way that the leading digit(s) gives the stem and the trailing digit gives the leaf.

Example: Set up the stem and leaf display for the data given below.

26, 28, 20, 20, 21, 22, 25, 25, 18, 25, 15, 20, 18, 20, 25, 25, 22, 35, 30, 31, 15, 20, 29.

Solution: The stem and leaf display of the given data is as follows:

Stem	Leaves
1	5 5 8 8
2	0 0 0 0 0 1 2 2 5 5 5 5 5 6 8 9
3	0 1 5

Example: Set up the stem and leaf display for the data given below.

5.35, 4.75, 4.30, 5.47, 4.85, 6.62, 3.54, 4.87, 6.26, 5.48, 7.27, 8.45, 6.05, 4.76, 5.91

Solution: Let us round the decimal in one place and arrange the data in ascending order as follows:

3.5, 4.3, 4.8, 4.8, 4.9, 4.9, 5.4, 5.5, 5.5, 5.9, 6.1, 6.3, 6.6, 7.3, 8.5

In this data, the smallest value 3.54 is rounded to 3.5. The stem of this value is 3 and leaf is 5.

The stem and leaf display is shown below.

Stem	Leaves
3	5
4	3 8 8 9 9
5	4 5 5 9
6	1 3 6
7	3
8	5

Exercise:

1. Draw a bar diagram to represent the following frequency distribution of the number of different surgical operations performed in a hospital in a given week.

Types of operation	General	Thoracic	Prostate	Urologic	Neuro
Number of cases	90	20	65	79	25

2. The medical institute has assigned following beds for different departments to provide health services to the people.

Department	**No. of bed**
Medicine	112
Surgery	128
Obs. /Gyn	72

Orthopedics 48
Others 115
Draw a simple bar diagram to represent the above information.

3. A survey was conducted to gather information on the purification methods practiced by households to treat and protect water. The findings of the survey are shown below.

Water purification/protection method	**% of households**
Boil	5.9
Use of purifying tablets	0.6
Filter	4.5
Cover storage vessel	60.2
Keep source clean	20.3
Others	34.4

Draw a simple bar diagram to represent above information.

4. Percentage of households reporting water-borne diseases in their communities in the last one month is given below.

Disease	Household percentage		
	Terai	Hills	Mountains
Diarrhea	12.9	15.8	31.6
Dysentery	5.3	7.2	18.1
Jaundice	3.3	2.1	3.9
Typhoid	3.3	2.5	4.9
Cholera	1.2	1.0	3.0

Represent the above information by multiple bar diagram.

5. Represent the following data of health personnel per 100000 populations and its growth between 1975 and 1994 by means of multiple bar diagram.

Personnels	Number of personnel		
	1974/75	1984/85	1993/94
Doctor	2.7	4.5	7.3
Nurse	6.5	15.9	22.3
Kaviraj/Vaidhya	1.3	1.7	2.6
Health assistant	6.4	10.2	21.8
Health worker	0.0	31.2	116.5

6. The following table shows morbidity and mortality trend of Japanese Encephalitis in Nepal(1998-2003)

Year	Cases	Deaths
1998	1161	149
1999	2924	434
2000	1729	169
2001	1888	275
2002	842	168
2003	330	69
Total	8874	1264

Represent the above data by multiple bar diagram.

7. Pathology department section wise total tests done in the first quarter of a year is given below

Tests	Baishak	Jestha	Ashad
Biochemistry	5520	5678	5784
Clinical Pathology	11590	14030	16290
Microbiology	1294	1511	1720

Represent the above data by a subdivided bar diagram.

8. Types of waste generation on Health care institutions in a valley are given below:

Types of waste	**Waste in kg/day**
Infectious	289.28
Non infectious	1516.19
Saline bottles	382.00
Total	2187.47

Represent the above information by pie chart.

9. Construct a pie chart for the following data:

Knowledge about Exclusive Breast Feeding	Number
Yes	112
No	14
Do not Know	39

10. Prepare sector diagram for the following data.

Types of leprosy	**No. of patients**
Tuberculoid	148
Lepromatous	64
Indeterminate	18
Border line	10
Total	240

11. In a survey, households were asked whether children surviving below 5 years of age were registered or not at the vital registration office. The table below presents the reasons given for not registering the birth of surviving children below 5 years of age.

Background characteristics	**No. of households**
Do not know	1614
No time	2105
No need	1057
Too far	122
Others	482

Draw a pie chart for the above information.

12. The table depicts the comparison of surgeries performed during the FY 02/03 and FY 01/02 in Himalayan Eye Hospital.

Surgeries	**01/02**	**02/03**
Cataract	2744	1463
Glaucoma	22	26
Vitreoretinal	16	58
Other intra ocular	104	100
Extra ocular	531	416
Total	3417	2063

Represent the information in pie chart.

13. The following data give the distribution of younger and older adults by fasting plasma glucose level.

Plasma glucose level (mg/dl)	Number of adults	
	Younger	Older
80-89	56	97
90-99	109	188
100-109	162	251
110-119	37	174
120 +	36	90
Total	400	800

Compare the two groups diagrammatically. Justify the diagram you have chosen.

14. Draw a histogram, frequency polygon for the following distribution of 727 deaths from Cholera (for various ages) classified according to age.

Age	0-5	5-10	10-15	15-20	20-25	25-30	30-45	45-70
No. of deaths	322	213	70	27	26	17	33	19

15. In a study of the association between smoking and coronary heart diseased, a team of clinical researchers obtained the following frequency distribution of systolic blood pressure reading (mm/Hg) for 37 smokers.

Systolic blood pressure	Frequency
89.5 – 109.5	5
109.5 – 129.5	15
129.5 – 149.5	10
149.5 – 169.5	3
169.5 – 189.5	2
189.5 – 209.5	2
Total	37

Represent the above information by histogram.

16. In a study of diabetic patients, the following data were obtained.

Age (in years)	No. of patients
10 – 20	2
20 – 30	10
30 – 40	25
40 – 50	15
50 – 60	5
60 – 70	3

Represent the above information by histogram and locate mode graphically.

17. The following is the set of birth weights of children with severe idiopathic respiratory disease syndrome measured to the nearest 100g.

Birth weight (kg)	No. of children
1.0 – 1.4	12
1.4 – 1.8	12
1.8 – 2.2	7
2.2 – 2.6	10
2.6 – 3.8	9

Draw a histogram to display this data.

18. The following data represents the age of the patients recorded in a hospital.

Age	5 – 14	15 – 24	25 – 34	35 – 44	45 – 54	55 – 64	65 - 74
Frequency	20	15	12	45	28	20	10

From the above data, draw less than and more than ogive on the same graph paper and estimate the median age.

19. The birth weight (in kilograms) of 15 children is recorded as: 2.0, 2.1, 2.3, 3.0, 3.1, 2.7, 2.8, 3.5, 3.1, 3.7, 4.0, 3.5, 4.1, 3.9 and 2.8. Construct a stem and leaf display.
20. The details of daily protein intake (gm) of 20 adult males are given below:

65 68 22 63 43 32 43 42 25 49 27 27
74 38 49 28 42 51 30 36

Construct a stem and leaf display.

21. The arm spans of 25 adult male employees of a certain firm were measured. They were rounded to the nearest centimeter and recorded.

177 173 183 178 186 168 172 178 171 186 172
179 183 179 169 180 170 186 180 181 179 184
174 181 187

Construct the stem and leaf display.

Measures of Central Tendency
It is a single value within the range of data, which represents a group of individual values in a simple and concise manner and concentrates in the middle of the distribution.

Objectives of central tendency
The objective of studying measures of central tendency is to get one single value that represents the entire data and to facilitate comparison.

Measures of central tendency
The various measures of central tendency are:

1. Arithmetic mean or simply mean
2. Median
3. Mode
4. Geometric mean
5. Harmonic mean

Arithmetic mean
The arithmetic mean is the most popular measure of central tendency. It is calculated by the ratio of sum of all the observation to the total number of observations.

Individual series
For a set of n observations $x_1, x_2, \ldots, x_n$ their arithmetic mean is denoted by $\bar{x}$ and is given by,

$$\bar{x} = \frac{x_1 + x_2 + \ldots\ldots + x_n}{n} = \frac{\sum x}{n}$$

Where, Σx = the sum of all observations, n = total number of observations.
This method of finding arithmetic mean is known as direct method.

Deviation method (Shortcut method)
The formula for calculating mean by this method is given by,

$$\bar{x} = a + \frac{\sum d}{n}$$

Where, a = assumed mean,
d = x – a, deviation from arbitrary value a (we call (x – a) as change in origin)
Σ d = sum of deviation taken from assumed mean
n = number of observations

Discrete series
Direct method
Let $x_1, x_2, x_3, \ldots\ldots, x_n$ are n variate values and $f_1, f_2, f_3, \ldots, f_n$ be their respective frequencies. Then their arithmetic mean is given by,

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \ldots\ldots + f_nx_n}{f_1 + f_2 + \ldots\ldots\ldots + f_n} = \frac{\sum fx}{\sum f} = \frac{\sum fx}{N}$$

Where, $N = f_1 + f_2 + \ldots + f_n$

Deviation method (Shortcut method)

The formula used for the computation of arithmetic mean for discrete series by this method is:

$$\bar{x} = a + \frac{\sum fd}{N}$$

where, a = assumed mean

d = x – a, deviation of items from their assumed mean

N = total frequency.

Continuous series

The same above two methods are also applied for the continuous series but x is taken as the mid value of the class interval.

Step deviation method

The formula used for the computation of arithmetic mean for continuous series by this method is:

$$\bar{x} = a + \frac{\sum fd'}{N} \times h$$

where, a = assumed mean

$$d' = \frac{x - a}{h}$$

h = magnitude of class interval

N = total frequency.

Note: Mean changes with the same amount when the value of a variate are increased or decreased by the certain amount.

Mathematical properties of arithmetic mean

Property 1 The algebraic sum of the deviations of the given set of observations from their arithmetic mean is zero.

Property 2 The sum of squares of deviation of the given set of observations is minimum when taken from the arithmetic mean.

Property 3 Mean of the combined series

If $n_1, n_2, \ldots\ldots, n_k$ are the sizes; $\overline{x_1}, \overline{x_2}, \ldots\ldots, \overline{x_x}$ be the means of kth series of observations, then the mean of the combined series of size $n_1 + n_2 + \ldots\ldots + n_k$ is

$$\overline{x_c} = \frac{n_1\overline{x_1} + n_2\overline{x_2} + \ldots\ldots + n_k\overline{x_k}}{n_1 + n_2 + \ldots\ldots + n_k}$$

Weighted arithmetic mean

Let $W_1, W_2, \ldots\ldots, W_n$ be the weights assigned to the variate values $x_1, x_2, \ldots, x_n$, then the formula for computing the weighted arithmetic mean denoted by $\bar{x}_w$ is defined by,

$$\bar{x}_w = \frac{W_1x_1 + W_2x_2 + \ldots\ldots + W_nx_n}{W_1 + W_2 + \ldots\ldots + W_n} = \frac{\sum Wx}{\sum W}$$

Note: 1. If the weights are equal, then $\bar{x} = \bar{x}_w$. If $W_1 = W_2 = \ldots = W_n = W$ say,

$$\bar{x}_w = \frac{W(x_1 + x_2 + ... + x_n)}{nW} = \frac{x_1 + x_2 + ... + x_n}{n} = \bar{x}$$

2. If greater weights are assigned to greater values and smaller weights are assigned to smaller values, then $\bar{x} < \bar{x}_w$
3. If greater weights are assigned to smaller values and smaller weights are assigned to greater values, then $\bar{x} > \bar{x}_w$

Median

The median is that value of the variable which divides the group in two equal parts, one part comprising all the values greater and the other, all the values less than median. The median depends in the position occupied by a value in the frequency distribution so it is also known as 'positional average'.

Calculation of median

Individual series: We first arrange the given set of observations into ascending or descending order. Then median is calculated by the relation,

$$\text{Median } (M_d) = \text{Size of } \left(\frac{n+1}{2}\right)^{th} \text{ item}$$

Where n is the number of observations.

On calculating the median in individual series, there arise two cases,

Case I: When the number of observations is odd, then median will be the middle most values after arranging the data.

Case II: When the number of observations is even, then there will be two middle most values. In such case, median is calculated by taking arithmetic mean of such two middle most values.

Discrete series: In discrete frequency distribution, the steps in calculating median are as follows:

1. Prepare less than cumulative frequency (c.f)
2. Find $\left(\frac{N+1}{2}\right)$
3. In cumulative frequency column, find the c.f just greater than or equal to $\left(\frac{N+1}{2}\right)$.
4. The corresponding value of the variable gives median.

Continuous series: The necessary steps for determining the value of the median in a continuous frequency distribution are:

1. Prepare less than cumulative frequency.
2. Find $\frac{N}{2}$, where $N = \Sigma f$
3. In the c.f. column, find the c.f. just greater than of equal to $\frac{N}{2}$.
4. The corresponding class with this c.f. contains median and the class is known as median class.

5. The exact value of the median is calculated by using the relationship,

$$\text{Median} = l + \frac{\frac{N}{2} - c.f}{f} \times h$$

Where l = lower limit of the median class,
f = frequency of the median class,
c.f = cumulative frequency preceding the median class,
h = magnitude of the median class

Partition values

The values which divide the series into a number of equal parts is called the partition values. Median may be regarded as a particular partition value which divides the given data into two equal parts.

Quartiles

The values which divides the data into four equal parts is known as quartiles. Obviously there will be three such points Q_1, Q_2 and Q_3 such that $Q_1 \le Q_2 \le Q_3$. Q_1 is known as lower or first quartile is the value which has 25% of the items of the distribution below it and 75% of the items are greater than it. Q_2 is the second quartile and Q_3 is the upper or third quartile has 75% of the observations below it and 25% of the observations above it.

Deciles

The values which divides the data into ten equal parts is known as deciles. Obviously there will be nine such points D_1, D_2 ,……D_9 such that $D_1 \le D_2 \le$……$\le D_9$.

Percentiles

The values which divides the data into hundred equal parts is known as percentiles. Obviously there will be ninety nine such points P_1, P_2 ,……P_{99} such that $P_1 \le P_2 \le$……$\le P_{99}$.

Calculation of partition values

Individual series

Quartiles, deciles and percentiles can be obtained by the following formula:

Q_i = size of $i\left(\frac{n+1}{4}\right)^{th}$ item, i = 1,2,3

D_i = size of $i\left(\frac{n+1}{10}\right)^{th}$ item, i = 1,2,3,….,9

P_i = size of $i\left(\frac{n+1}{100}\right)^{th}$ item, i = 1,2,3,…..,99

Discrete series

Steps in calculating the partition values are as follows:

1. Prepare less than cumulative frequency (c.f)
2. Find $i\left(\frac{N+1}{4}\right)^{th}$ item, i = 1,2,3

   $i\left(\frac{N+1}{10}\right)^{th}$ item, i = 1, 2, 3,…., 9

$i\left(\frac{N+1}{100}\right)^{th}$ item, i = 1, 2, 3,….., 99

For quartile, decile and percentile resp. Where N = Σ f = total frequency

3. In cumulative frequency column, find the c.f. just greater than or equal to $i\left(\frac{N+1}{4}\right), i\left(\frac{N+1}{10}\right), i\left(\frac{N+1}{100}\right)$ for quartiles, deciles and percentiles respectively.
4. The corresponding value in the variable gives the desired partition values.

Continuous series

The necessary steps for determining the value of the partition values in a continuous frequency distribution are:

1. Prepare less than cumulative frequency.
2. Find $i\frac{N}{4}, i\frac{N}{10}, i\frac{N}{100}$ for quartiles, deciles and percentiles respectively.
3. In the c.f column, find the c.f just greater than or equal to $i\frac{N}{4}, i\frac{N}{10}, i\frac{N}{100}$.
4. The corresponding class with this c.f contains desired partition values and the class is known as partition class which is named as quartile class, decile class and percentile class.
5. The exact value of the quartiles, deciles and percentiles is calculated by using the formula,

$$Q_i = l + \frac{i\frac{N}{4} - c.f}{f} \times h \; ; \; i = 1, 2, 3$$

$$D_i = l + \frac{i\frac{N}{10} - c.f}{f} \times h \; ; \; i = 1, 2, 3, \ldots, 9$$

$$P_i = l + \frac{i\frac{N}{100} - c.f}{f} \times h \; ; \; i = 1, 2, \ldots, 99$$

Where l = lower limit of partition value class,
f = frequency of the partition value class,
c.f = cumulative frequency preceding the partition value class,
h = magnitude of the partition value class

Mode

The variate value which occurs most frequently in the given set of data is known as mode. The series may have no mode, one mode (unimodal), having two modes (bimodal) and more than two modes (multimodal).

Calculation of mode

Individual series

It is easy to locate mode in case of individual series. Here the number of times of repetition of each variate values is noted, the value which occurs maximum number of times is the modal

Discrete series

In discrete series, the variate value having the maximum frequency is the mode.

Continuous series

In case of continuous frequency distribution, mode is calculated in the following manner.

1. We locate the modal class, i.e. the class having maximum frequency.
2. The value of mode is calculated by the relation,

$$\text{Mode } (M_o) = l + \frac{\Delta_1}{\Delta_1 + \Delta_2} \times h$$

Where l = lower limit of the modal class,

$\Delta_1 = f_1 - f_0$, $\Delta_2 = f_1 - f_2$

f_1 = frequency of the modal class

f_0 = frequency of the class preceding the modal class,

f_2 = frequency of the class succeeding the modal class

h = magnitude of the modal class

While calculating mode in continuous series, the class interval must be of equal magnitude. If not they must be made equal.

Calculation of mode by grouping method

We use grouping method to calculate mode in the following cases:

i. if the maximum frequency is repeated,
ii. if there is an irregular distribution of the data,
iii. if the maximum frequency occurs at the very beginning or at very end,
iv. if the pressure difference between the maximum frequency and the frequency preceding or succeeding is very small.

Grouping method is divided into two parts. The first part consists of grouping table and the second one is the analysis table. Grouping table consists of six columns. In these six columns, follow the procedure as stated below:

i. Column I: Write down the original frequencies.
ii. Column II: Add the frequencies pair wise (2 by 2) and place them in the mid position of such two frequencies.
iii. Column III: Leave the first frequency and add rest of the frequencies pair wise.
iv. Column IV: Add the frequencies 3 by 3.
v. Column V: Leave the first frequency and add rest of the frequencies 3 by 3.
vi. Column VI: Leave the first two frequencies and add rest of the frequencies 3 by 3.

Analysis table

i. The variate values are written in the column and the six columns in row.
ii. In the body of the table give value 1 to each variate values which have been summed up in consisting the maximum frequency.
iii. Add all the 1's of each column.
iv. The variate value which has the maximum column total gives the mode in discrete series. In continuous series, the maximum column total gives the modal class and mode is calculated by the relation given above.

Remarks:

1. If $f_0 > f_1$, then mode is calculated by the relation,

Mode (M_o) = $l + \frac{f_2}{f_0 + f_2} \times h$

2. If $\Delta_1 - \Delta_2 = 0$, then mode is calculate by the relation,

   $M_o = l + \frac{\Delta_1}{|\Delta_1| + |\Delta_2|} \times h$

   Where, $|\Delta_1|$ = absolute value of Δ_1, $|\Delta_2|$ = absolute value of Δ_2
3. For bimodal or multimodal distribution, mode is calculated by the empirical relation, Mode = 3 median – 2 mean

Geometric mean

In a set of n non zero and non negative observations, the geometric mean denoted by GM or simply G is defined as the n^{th} root of their product.

Thus for $x_1, x_2, \ldots, x_n$ observations, then their geometric mean is given by,

$G = \sqrt[n]{x_1 x_2 .....x_n} = (x_1.x_2 .....x_n)^{\frac{1}{n}}$ ……..(1)

Formula to calculate geometric mean in individual series,

$G = \text{antilog}\left(\frac{1}{n}\sum \log x\right)$

In discrete series, if $f_1, f_2, \ldots, f_n$ be the corresponding frequencies of n non negative and non zero variate values $x_1.x_2 \ldots\ldots x_n$ respectively, then their GM is given by,

$G = \text{antilog}\left(\frac{1}{N}\sum f.\log x\right)$

In case of continuous series, we take the value of x as the mid value of the class interval.

Problems on measures of central tendency

1. A sample of 10 persons showed the number of attacks of cold per person as follows: 7, 2,5,0,1,5,5,6,7,6. Calculate mean, median and mode. **(4.4, 5, 5)**
2. In a random sample of 10 patients waiting in a clinic to be served, the waiting time in minutes was observed as: 20, 15, 30, 17, 50, 12, 13, 15, 18 and 10. What is the average waiting time for the patients in the clinic? **(20 min)**
3. Compute the mean, median and mode from the following figures representing the weight (in pounds) of 10 newborn babies at ABC hospital on a certain day: 8,6,7,7,7,5,9,7,8,6 **(7,7,7)**
4. The disintegration time in minutes of different batches of 10 coated tablets is: 22,25,21,22,17,18,21,20,20,19. Calculate the median time. **(20.5)**
5. The systolic blood pressure of the first 11 persons is recorded as follows:
   120 118 152 160 150 134 125 145 135 139 160
   Calculate the median and mode. **(139,160)**
6. Find the mean days of confinement after delivery in the following series.

Days of confinement	6	7	8	9	10
No. of patients	5	4	4	3	2

(7.61)

7. For a certain frequency table which has only been partly reproduced here. The mean was found to be 1.46. Calculate the missing frequencies if 200 days are observed.

No. of accidents	0	1	2	3	4	5
Frequency(No. of days)	46	?	?	25	10	5

(38,76)

8. The following table shows the age distribution of a certain disease reported during a year in a particular state.

Age	5 – 10	10 – 15	15 – 20	20 – 25	25 – 30	30 – 35
No. of cases	5	10	120	22	13	5

Calculate the average number of cases reported in a year. **(18.73)**

9. In a city general hospital survey conducted to know the no. of day's patient stayed in a hospital following an operation. The data are given below.

Hospital stay (in days)	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24
No. of patients	32	108	67	28	14	7	3	1

Calculate the average stay days. **(7.1 days)**

10. In a survey, mothers with children less than 5 years of age were asked about the health of their children. Distribution of children who suffered from cough and cold in past two weeks and that were taken to a health care provider for treatment are given below.

Age (in months)	No. of children
< 6	149
6 – 11	269
12 – 23	362
24 – 35	374
36 – 47	373
48 – 59	345

Calculate the median age in months who suffered from cough and colds. **(28.5)**

11. A data on the situation of HIV/AIDS in Nepal by age group in Nov 30, 2005 is given below.

Age group	No. of persons
0 – 4	66
5 – 9	59
10 – 14	33
15 – 19	358
20 – 24	114
25 – 29	1426
30 – 39	1991
40 – 49	510
50 above	90

Calculate the partition values Q_1, Q_3, D_6, P_{10}, P_{35}, P_{75} who suffered from HIV/AIDS.

12. The mean weight of 32 students in the class is 59.5 kg. The mean weight of boys in the class is 64 kg and that of girls is 54.4 kg. Find the number of boys and girls in the class.

(17,15)

13. A candidate obtained 90, 82, 93, 76 marks out of 100 in Biostatistics in the 1st term, 2nd term, 3rd term and assessment. If 5%, 5%, 80%, 10% weights are allotted respectively in each exam, find the candidates weighted arithmetic mean. **(90.6)**
14. In a connection with certain requirement the weights attached to the academic performances in different levels Certificate level, B. P.H, M. P.H was 2:3:5. If the

candidate has secured 65%, 60% and 75% respectively in the above examinations, find the weighted average of his performance. **(68.5%)**

15. An examination was held to decide the award of a scholarship. The weights assigned to various subjects were different. The marks obtained by three candidates are given below:

Subjects	Weights	Marks obtained by		
		A	B	C
MTH 191	3	63	50	65
CHEM 172	2	65	85	78
MLS 101	3	60	60	52
MBL 102	1	70	80	88

If the candidate getting the highest marks is to be awarded the scholarship, who should get it? **(C)**

16. The mean weight of a group of 26 students in B. P.H was calculated by 65kg. It was later discovered that one weight was misread as 65kg instead of 56kg. Calculate the correct mean weight of B. P.H students. **(64.65)**
17. The mean salary paid to 1,000 employers of an establishment was found to be Rs.1804.0. Later on, after disbursement of salary, it was discovered that the salary of two employees was wrongly entered as Rs. 2970 and1650. Their correct salaries were Rs.1970 and 1850. Find the correct arithmetic mean. **(Rs. 1803.2)**
18. Find the value of median of the following distribution. **(37.61)**

Value (less than)	10	20	30	40	50	60	70	80
Frequency	5	15	35	81	90	118	130	140

19. Calculate mode from the following data. **(67)**

Height of person (in inches)	59	61	63	65	67	69	71
No. of persons	3	6	12	45	120	99	45

20. Calculate mode from the following data. **(15)**

X:	10	11	12	13	14	15	16	17	18	19
f:	6	9	4	2	10	8	7	5	1	3

21. Calculate mode from the following data. **(35)**

X:	30	31	32	33	34	35	36	37	38	39
f:	5	7	6	11	21	23	25	7	3	2

22. Calculate mode from the following frequency distribution. **(44.76**)

Class	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	100-110
f	4	2	18	22	21	19	10	3	1

23. Calculate geometric mean of:
    a) 1, 7, 8, 5, 12, 15
    b)

Year	1	2	3	4	5
Growth factor	1.07	1.08	1.10	1.12	1.18

    c)

X	2	3	4	5	6	7	8
f	2	4	7	4	4	2	1

    d)

Class interval	5-10	10-15	15-20	20-25	25-30
Frequency	3	5	12	6	2

Measures of dispersion

Measures of dispersion is a descriptive statistical measure used to measure the variation or spread or scatterness in a data set. It is also known as measures of variation. It gives additional information that enables to judge the reliability of measure of central tendency. If all the values of a data set are identical, there is no dispersion. If they are not identical, there is dispersion. The amount of dispersion of different values may be small, if the values are close together.

Absolute and relative measure of dispersion

A measure of dispersion is said to be an absolute if it is expressed in terms of original units of the data. If two statistical measures in same units are calculated there is no problem to compare these measures by calculating only absolute measures. But if two statistical measures with different measurements units are to be compared, it is difficult to compare because of different measurement units used. In such case we have to use the relative measure. A relative measure of dispersion is said to be relative measure if it is independent of units of the data. It is simply a coefficient.

Measures of dispersion

Following are common measures of dispersion

1. Range
2. Quartile deviation
3. Mean deviation
4. Standard deviation.

Range

Range may be defined as the difference between the highest and lowest measurement in a group of data. Let L be the largest or highest value and S be the smallest or lowest value. The range is denoted by R and computed as follows:

Range = L – S

Range is an absolute measure of dispersion. If one is intersested to compare the variability of data sets having different units of measurement, coefficient of range will be used which is calculated as follows.

$$\text{Coefficient of range} = \frac{L - S}{L + S}$$

Quartile deviation

This method measures the distance between quartiles. Since this method is based on two quartiles, it is also known as dispersion measured with quartiles.

Interquartile range = $Q_3 - Q_1$

Quartile deviation or semi interquartile range is defined as,

$$\text{Quartile deviation} = \frac{Q_3 - Q_1}{2}$$

Where Q_1 = first quartile and Q_3 = third quartile

The relative measure of quartile deviation is also known as coefficient of quartile deviation and is calculated as follows.

$$\text{Coefficient of quartile deviation} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

Quartile deviation is the most appropriate method in case of open class interval.

Mean deviation

The mean deviation or average deviation or mean absolute deviation is able to measurer the deviations from arithmetic mean or median or mode.

Individual series,

Mean deviation taken from A = $\frac{1}{n}\sum|X - A|$

Where A = mean or median or mode

Mean deviation taken from mean = $\frac{1}{n}\sum|X - \bar{X}|$

Discrete and continuous series

Mean deviation taken from A = $\frac{1}{N}\sum f\,|X - A|$

Where A = mean or median or mode

In case of continuous series X is taken as mid value of the class interval.

The relative measure of mean deviation is known as coefficient of mean deviation from mean or median or mode and is computed as,

Coefficient of mean deviation from A = $\frac{\text{Mean deviation from A}}{A}$

It ignores the negative sign and can not be calculated exactly in case of open class interval.

Standard deviation

Standard deviation is the positive square root of the average of the square of the deviations of the measurements from their mean. It is denoted by σ.

Individual series

$$\sigma = \sqrt{\frac{1}{n}\sum(X - \bar{X})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

On further simplification (1) can be written as,

$$\sigma = \sqrt{\frac{\sum X^2}{n} - \left(\frac{\sum X}{n}\right)^2}$$

Discrete and continuous series

$$\sigma = \sqrt{\frac{1}{N}\sum f\,(X - \bar{X})^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots (2)$$

On further simplification (2) can be written as,

$$\sigma = \sqrt{\frac{\sum fX^2}{N} - \left(\frac{\sum fX}{N}\right)^2}$$

But X is taken as the mid value of the class interval incase of continuous series.

Square of standard deviation is known as variance and is denoted by σ^2.

Coefficient of variation (C.V)

It is a relative measure of dispersion based on standard deviation. In order to compare the variability between two sets of data, coefficient of variation can be used as a useful method.
C.V is defined as,

$$C.V = \frac{\sigma}{\bar{X}} \times 100$$

It is a number expressed in percentage.

Comparing variability

For comparing the variability between two or more than two sets of data, a distribution having more coefficient of dispersion is considered as more variable or more heterogeneous or less consistent or less uniform or less equitable or less stable etc.

Exercise

1. The length of hospital stay in days for six patients is observed. The values are 10, 5, 2, 3, 2, 6. Calculate quartile deviation and its relative measure.

2. Find the SD of incubation period of 9 patients where it was found to be 14, 13, 11, 15, 10, 7, 9, 12 and 10.

3. A certain drug was prescribed for 6 patients complaining of insomnia. The number of hours slept after drug is administered is 6.8, 6.9, 17.8, 6.4, 5.8, 3.0. Calculate standard deviation of the given data. **(4.45)**
4. Calculate range and its coefficient from the following data:
   Class 25-30 30-35 35-40 40-45 45-50
   Freq 3 5 10 12 6 **(25, 0.33)**
5. Calculate QD and its coefficient from the following data **(4, 0.061)**

Height (in inches)	60	62	64	66	68	70	72	74	76
No. of persons	15	25	10	18	30	16	8	7	5

6. Find the most appropriate measure of dispersion for the age distribution of TB defaulter in a certain city. **(8.06)**

Age	less than 20	20 – 30	30 – 40	40 – 50	more than 50
No. of persons	1	2	4	7	5

7. Calculate semi interquartile range and its coefficient from the data given below:

Count range	60-65	66-71	72-77	78-83	84-89	90-95
Occurrence of particle	12	7	7	7	4	5

8. You are given two variables A and B, using quartile deviation state which is more variable?

A		B	
Mid point	Frequency	Mid point	Frequency
15	15	100	340
20	33	150	492
25	56	200	890
30	103	250	1420
35	40	300	620

40	32	350	360
45	10	400	187
		450	140

9. Given below are the weights of 9 students. Calculate mean deviation from the mean as well as median.
   Weight (in kg): 40, 45, 42, 51, 47, 50, 55, 54, 57 **(4.89, 4.78)**
10. Calculate mean deviation from mean, median and mode from the following frequency distribution.

Age in years	6	7	8	9	10	11	12
No. of boys	3	6	9	13	8	5	7

(1.39, 1.35, 1.35)

11. Calculate mean deviation from mean and median from the following data.

Marks	12	16	20	24	28	32	36	40
No. of students	1	14	25	27	18	9	4	2

(4.48, 4.48)

12. Find out mean deviation from median and its coefficient from the following data.

Class interval	0 – 10	10 – 20	20 – 30	30 – 40	40 – 50
Frequency	5	7	4	8	10

(12.87, 0.41)

13. Calculate mean deviation from mean, median and mode from the following grouped frequency distribution.

Class interval	0 – 4	4 – 8	8 – 12	12 – 16	16 – 20
Frequency	4	6	8	5	2

(3.84, 3.83, 3.76)

14. The following data presents the retail price of a equipment in 10 different shops. Calculate standard deviation of the prices. **(0.122)**

Price (Rs. Lakh)	2.00-2.10	2.10-2.20	2.20-2.30	2.30-2.40	2.40-2.50
No. of shops	2	0	4	3	1

15. Find the standard deviation of the following data: **(19.76)**

Age under	10	20	30	40	50	60	70	80
No of persons dying	15	30	53	75	100	110	115	125

16. Calculate SD from the following data: **(14.6)**

Temp (°C)	No. of days	Temp (°C)	No. of days
-40 to -30	10	0 to 10	65
-30 to -20	24	10 to 20	180
-20 to -10	30	20 to 30	14
-10 to 0	42		

17. Calculate standard deviation for the following distribution of 727 deaths from Cholera (for various ages) classified according to age. **(11.66)**

Age (in years)	0–5	5–10	10–15	15–20	20–30	30–45	45–70
No. of deaths	322	213	70	27	43	33	19

18. From the following frequency distribution of weights (lbs) for 100 persons calculate the standard deviation and coefficient of variation.

Weight (lbs)	No. of persons
130.5 –140.5	10
140.5 –150.5	20
150.5 –160.5	30
160.5 –170.5	20
170.5 –180.5	10
180.5 –190.5	10

19. The recommended daily allowance of calcium for adults is 800mg. A nutritionist suspects that people living in a particular area have a low intake of calcium. To test out her suspicion she measures the calcium intake of 50 people from the area. The data is recorded rounded down to the nearest mg.

Intake (mg)	No. of adult
0 – 200	1
200 – 400	1
400 – 600	12
600 – 800	16
800 – 1000	12
1000 – 1200	7
1200 – 1400	1
Total	50

Compute standard deviation, and interquartile range for this sample.

20. Diarrhoeal diseases in Nepal remain one of the major public health problems especially among children under 5 years of age. The number of children with blood in stool suffering from diarrhea is given below

Age (in months)	No. of children
< 6	64
6 – 11	151
12 – 23	236
24 – 35	242
36 – 47	187
48 – 59	145

Calculate the appropriate measure of dispersion.

21. An analysis of the monthly wages paid to workers in firm A and B belonging to the same industry gives the following results.

	Firm A	Firm B
No. of workers	500	600
Average monthly wages (Rs)	4800	4650
Variance of the distribution of wages	4000	6250

(i) Which firm pays a larger wage bill?
(ii) In which firm there is greater variability in individual wages?
(iii) Find the combined mean wages of the two firms taken together. **(B, B, Rs. 4718.18)**

22. The erythrocyte sedimentation rate (mm/hour) of 15 male and 10 female TB patients before start of the treatment is given below.

	Male	Female
Mean	88.73	94.90
Standard Deviation	25.32	15.76

Find which character shows greater variation. **(Male)**

23. A training company is considering employing one of two training programs. Two groups were trained for the same task. Group I was trained by program A, group II by program B. For the first group, the time required to train the employees had an average of 32.11 hours and variance 68.09. In the second group, the average was 19.75 hours and the variance was 71.14. Which training program has less relative variability in performance? **(Program A)**

24. The following data give the ages and weights of 5 persons in a locality. Show whether the variation in ages is more than the variation in weights.

Ages (years)	32	45	35	40	45
Weights (kg)	60	65	65	70	72

(Age)

25. Suppose two groups of human males yield the following information.

	Group A	Group B
Age	24 years	15 years
Mean weight	145 lbs	80 lbs
Variance	100 lbs	100 lbs

Find which is more variable the weights of 24 years olds or the weights of 15 years old. **(15 years old)**