Dept. of Public Health Dentistry

Biostatistics

By. Dr. Kajal Patel

INTRODUCTION:-

STATISTICS is defined as the science of collecting, summarizing, presenting, analyzing and interpreting the data.

STATISTICS means a measured or counted fact or piece of information state as a figure such as height of one person, birth of a baby, etc.

They are collected from experiments, records and surveys.

Bio-Statistics

BIOSTATISTIC is that branch of statistics concerned with mathematical facts and data related to biological events.

Medical statistics go under different names when applied in different fields such as;

1) Health statistics in public health or community health.

2) Medical statistics in medicine related to the study of defect, injury, disease, efficacy of drug, serum and line of treatment, etc.

3) Vital statistics in demography pertaining to vital events of births, marriages and deaths. These terms are overlapping and not exclusive of each other.

Principles of Biostatistics:

- Collection of data
- Presentation of data
- Summarization of data
- Analysis of data
- Interpretation of data

EPIDEMIOLOGY and BIO-STATISTICS

EPIDEMIOLOGY and BIO-STATISTICS are sister sciences or disciplines.

Epidemiology collects facts relating to groups of population in places, times and situations

Bio-statistics converts all facts into figures and at the end translates them into facts, interpreting the significance of their results.

Both the science of epidemiology and bio-statistics deal with

facts → figures → facts , which is termed as quantitative methodology.

USE OF BIOSTATISTICS

1. To identify causes of disease.
2. Describe the incidence and prevalence of diseases
3. To identify which area rural or urban –are more or less affected by the diseases (difference between two populations)
4. Evaluate standard of health.
5. Help for planning of health programs
6. Evaluate the measures adopted.
7. To fix priorities in public health programs.

Application of statistics: As a science

- To define what is normal or healthy in a population and to find out the limits of normality in variables
- To find the difference between means and proportions of two places or in different periods
- To find the action of drug
- To compare the action of two different drugs or two successive dosages of the same drug
- To find the relative potency of a new drug with respect to a standard drug
- To find out action/efficacy of a drug, operation or line of treatment
- To find correlation between two variables
- To identify signs and symptoms of a disease.

As a figures:

- What are the various causes of dental caries? Or PD disease?
- What are the leading causes of dental caries or PD disease?
- Whether a particular disease is rising or falling in severity and prevalence.
- Which age group, sex, social class of people, profession, place etc are affected the most?
- Which health programme should be given priority and what will be the requirements for the same?

Common statistical terms:

- DATA : A collective recording of observation either numerical or otherwise is called DATA.
- VARIABLE : A certain observation is made of a characteristic which varies from one person to the other is called a variable.
- OBSERVATION : An event and its measurements such as blood pressure (event) and 120 mm of Hg (measurement).
- POPULATION: The group of all individuals who are the focus of the investigation is known as population.

TYPES

Two types

1) Qualitative or discrete data-

When the data is collected on the basis attributed or qualities like sex, malocclusions, cavity etc. , it is called qualitative data.

Qualitative data are discrete in nature such as number of oral cancer in different years, population of different towns, persons with different food habits and so on.

2) Quantitative or continuous data-

When the data is collected through measurement using calipers, like arch length, arch width, fluoride concentration in water supply etc., it is called quantitative data.

SAMPLING

- SAMPLE : Group of individuals who are actually available for the investigation.
- POPULATION: The group of all individuals who are the focus of the investigation is known as population.
- TARGET POPULATION: Population of interest
- SAMPLING UNIT : The individuals entities that form the focus of the study
- SAMPLING FRAME : The list of sampling units

SAMPLING

- The sample is a portion of the population selected from a population in some manner.
- Two main objectives of sampling are :

  1.estimation of population parameters (characteristic)

  2. To test the hypothesis about the population from which the sample or samples are drawn.

- Application of sampling in community dentistry:

  1. Evaluation of oral health status of a community

  2. Evaluation of health education on oral hygiene.

  3. Studies on administrative aspects of the services like availability and utilization of oral health facilities in the community.

Ways of Sample selection

- 1] Purposive selection:
  - Purposively select the individuals who seem to represent the population.
  - Easy to carry out
  - Does not need the preparation of sampling frame.
  - Under-represent the rates of the population.

- 2] Random selection :
  - Does not mean haphazard.
  - But it indicates the chance o f the population unit being selected in the sample.

SAMPLING METHODS

1} NON PROBABILITY SAMPLING ..not truly reprentative

1) Quota sampling
2) Purposive or Judgment sampling
3) Network or snowball sampling
4) Convenience or Accidental sampling

2} PROBABILITY SAMPLING

5) Simple random sampling
6) Systematic sampling
7) Stratified random sampling
8) Cluster sampling

3} Other sampling methods

9) Multiphase sampling
10) Multistage sampling

SAMPLING METHODS

1.Quota sampling.

Right number of people are found to fill quotas.

some percentages.

2.Purposive or judgment sampling.

specific group or disease

3. Network or snowball sampling

interview few subjects with same criteria- they find others with same criteria-till get satisfactory sample.

4. Convenience or Accidental

examine the people you are able to contact… volunteers …inexpensive and less time

- 5. Simple Random Sampling
  - Each and every unit in the population has an equal chance o f being included in the sample.
  - Methods for simple random sampling are Lottery method and Table of random numbers.
- 6.Systemic Random Sampling
  - It is formed by Selecting one unit at random and then selecting additional units at evenly spaced interval till the sample of required size has been formed form it.
  - This method is used when a complete list of population is available.
  - First decide the sample interval. Sample interval (k)= no. of unit in the population/ size of sample.
  - One number is selected; add the sample interval in that unit to get another sample unit.

- 7. Stratified Random Sampling :
  - Here the population to be sampled is subdivided into groups known as strata, such that each group is homogeneous in its characteristic e.g. primary division into males and females.
  - A simple random sample is then chosen from each stratum e.g. secondary division of each of these categories into five age groups.
  - This type of sampling is used when the population is heterogeneous with regard to the characteristic under study.
  - Advantage: provides greater accuracy and can concentrate on wider geographical area.
  - Disadvantage: require more skill, time.

- 8. Cluster Sampling
  - This method is used when the population forms natural groups or clusters, such as , villages, wards blocks or children of school etc.
  - Simpler and require less time and cost
  - Higher standard error.
- 9. Multiphase Sampling ;
  - In this method, a part of the information is collected from the whole sample and a part from the sub-sample.
  - For example, in a school health survey, all children in the school may be examined-ones with oral health problems may be selected in the second phase-ones needing treatment may be selected in the third phase.

10. Multistage sampling :

- Sampling procedures carried out in several stages using random sampling techniques.
- Country – state – district – villages – household
- Employed in large country survey.

Sample size

- Bigger the sample, higher will be the precision of the estimates of the sample.
- The size of the sample is decided according to ,
  - (1) Characteristics of the sample
  - (2) Availability of experimental material, resources and other practical consideration.

Errors In Sampling

- There are two types of errors that arise in sampling investigation, viz., sampling error and non-sampling error.
- The sampling errors are errors that creep in due to the sampling process and could arise because of
- [I] faulty sampling design or [ii] small size of the sample.
- The non-sampling errors arise due to,

1) Coverage error – due to non-response or non-cooperation of the informant

2) Observational error – due to interviewers bias or imperfect experimental technique or interaction of both.

3) Processing error – due to errors in statistical analysis.

DATA

SOURCES:- Sources for collection of data are

1) **Experiments:-The data collected with specific objective by one or more workers are compiled and analyzed.**

2) **Surveys: -Surveys are carried out by trained team to find the incidence or prevalence of heath or diseases situations in a community such as incidence of dental caries or prevalence of periodontal disease. They are also made use of in operational research, such as assessment of existing condition and to study the merits of different methods adopted to control a disease.**

3) **Records:-Records are maintained as a routine in registers or books over a long period of time for various purposes such as for vital statistics –births, marriages and deaths.**

Data can be collected through either

a) **Primary source**:- Here the data is obtained by the investigator himself. This is first hand information.

- Primary data can be obtained using any one of the following methods
- -- Interview. Two types : personal and telephone

↓

Direct (structured and unstructured) and Indirect

--Oral health examination

--Questionnaire method

b) **Secondary source**:- The data already recorded is utilized to serve the purpose of the objective of the study e.g. the records of the OPD of dental clinics.

METHODS OF PRESENTATION

There are two main methods of presenting data;

1) Informal presentation
2) Formal presentation
   - A) Tabulation
   - B) Drawing
     - –Graphical presentation
     - –Diagrammatical presentation

TABULATION:

- Tabulation are device for presenting data from a mass of statistical data.
- Table can be simple or complex depending upon measurement of single set of items or multiple sets of items.
- A most common way of presenting data in the tables is known as frequency distribution table.

1. MASTER TABLE:

   Contain all data

2. SIMPLE TABLE:

   Present one characteristic of data

3. FREQUENCY DISTRIBUTION TABLE (quantitative data)

   Two column- classes (group) and frequencies

age group	student with carious teeth
9-11 yrs	12
12-14 yrs	23
15-17 yrs	31

- The data of variable characteristics are continuous such as height, weight, pulse rate, bleeding time, etc.
- they have range from the lowest to the highest.
- this range is divided into subgroups called classes. The class limits are the lowest and highest values that can be included in the class. For instance. In the class 5-14, 5 is the lower limit and 14 is upper limit.
- The difference between the upper and lower limit of a class is known as class interval of the class. Example, in the class 5-14, class interval = 14 -5 =9.

Basic rules while forming a frequency distribution table,

- Every table should contain a title as to what is depicted in the table.
- The number of class intervals should not be too many or too less. It may be preferably between 5 and 20. However, there is no rigidity about it.
- The class intervals should be at equal width.
- The class limits should be clearly defined to avoid ambiguity.

  For e.g., 0-4,5-9,10-14. etc.
- Each row and column should be clearly defined with the headings for each row and column.
- Units of measurement should be specified.
- If the data is not original, the source of the data should be mentioned at the bottom of the table.

DRAWING

Two kinds of drawings- graph and diagrams

- Diagram and graphs are one of the most convincing and appealing ways of depicting statistical results.

- Diagram and graphs are extremely useful because they are attractive to the eyes, give a bird's eye view to the entire data, have a lasting impression on the mind of the layman and they facilitate comparison of data relating to different time periods and regions.

Rules for construction of diagrams and graphs:

- Every diagram must be given a title that is self-explanatory.
- It should be simple and consistent with the data.
- Usually, the values of the variables are presented on the horizontal or X-axis and the frequency on the vertical line or Y-axis.
- The number of lines drawn in any graph should not be many so that the diagram does not look clumsy.
- The scale of presentation for the X and Y-axes should be mentioned at the right hand top corner of the graph.
- The scale of division of the two axes should be proportional and the divisions should be marked along with the details of the variables and frequencies presented on the axes.

Types of diagrams:

- Depending on the type of data, whether it is quantitative or qualitative, any one of the following diagrams may be chosen.

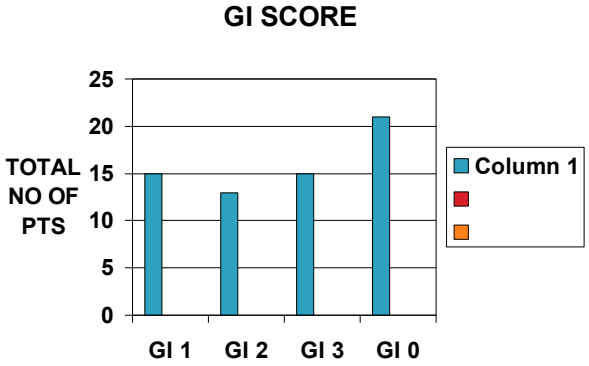
- For the presentation of the qualitative, discrete or counted data following diagrams are used,

1. Bar diagram
2. Pie or sector diagram
3. Pictogram or picture diagram
4. Cartogram or spot map.

- For the presentation of quantitative/continuous/ measured data following graphs are used

1. Histogram
2. Frequency polygon
3. Frequency curve
4. Line chart or graph
5. Cumulative frequency diagram
6. Scatter or dot diagram

1) BAR DIAGRAM:



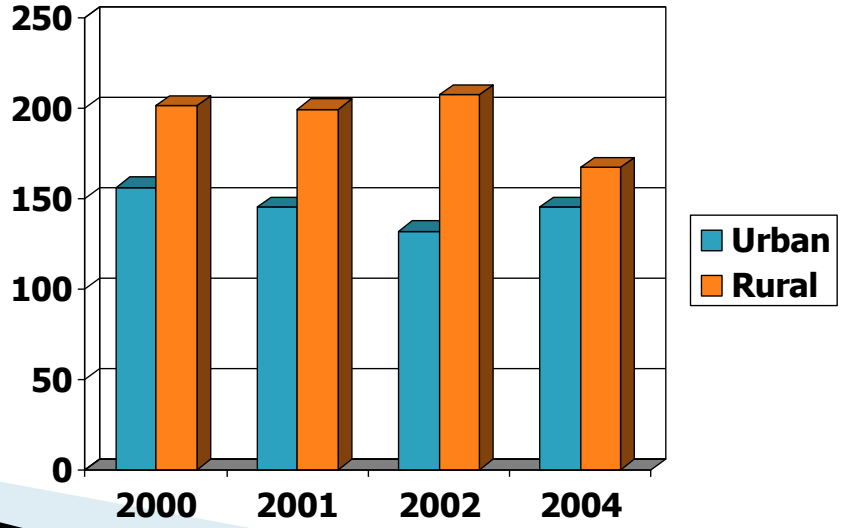
- Used to represent qualitative data.
- Represent only one variable.

  e.g. the number of people with D,M,F teeth in a particular age group may be shown by a bar diagram.
- The width of the bar remains the same and only the length varies according to the frequency in each category.
- The bars can be either vertical or horizontal.

1-a) MULTIPLE BAR

- Used to compare qualitative data with respect to a single variable (sex,age or region)
- Similar to bar diagram except that for each category of the variable a set of bars of the same width corresponding to the different sections without any gap in between the width and the length corresponds to the frequency
- Prevalence of periodontal disease in urban and rural area:

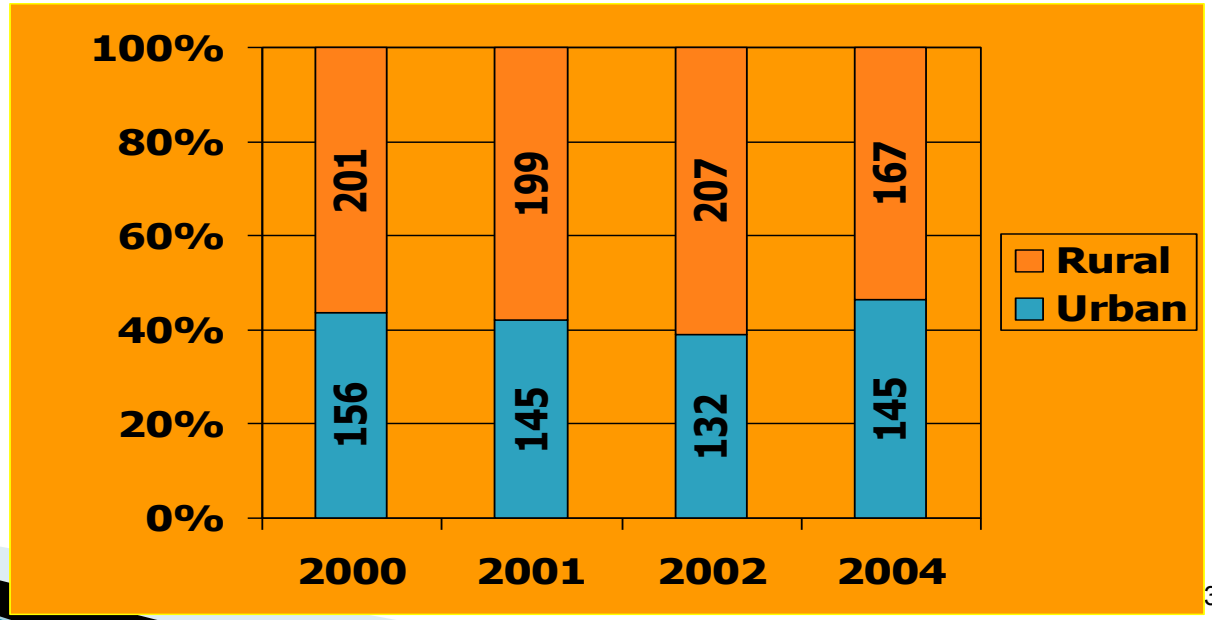
year	Urban	Rural
2000	156	201
2001	145	199
2002	132	207
2004	145	167



1-b) PROPORTIONAL BAR DIAGRAM :

- Used to represent qualitative data.
- To compare only the proportion of subgroups between different major groups of observation.
- Bars are drawn for each group with the same length; either as 1 or 100%. These are then divided according to the sub group proportion in each major group.
- Prevalence of periodontal disease in urban and rural area:

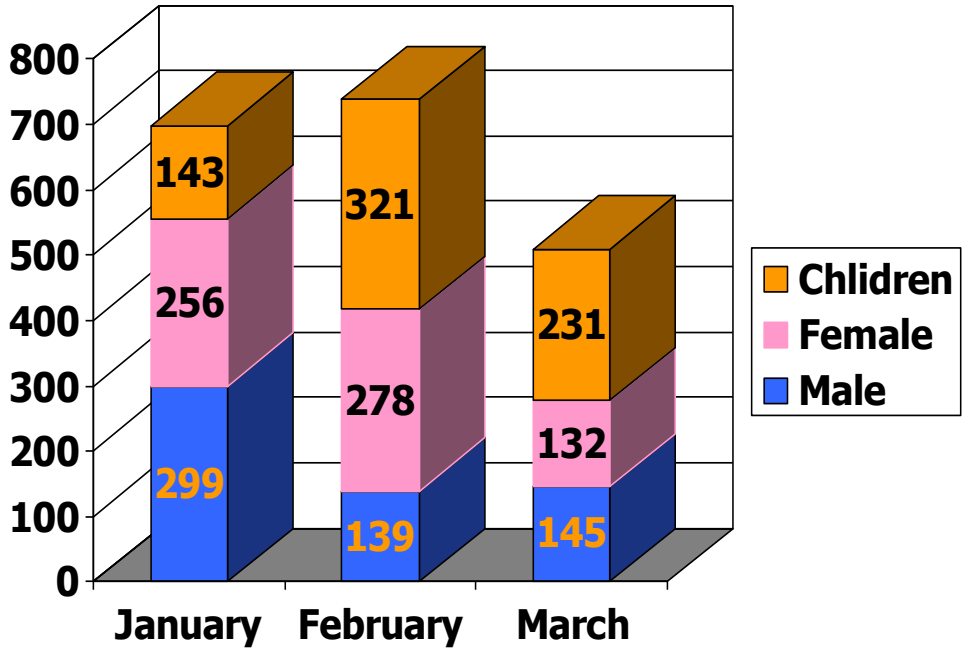
year	Urban	Rural
2000	156	201
2001	145	199
2002	132	207
2004	145	167



1-c) COMPONENT BAR DIAGRAM

- When it is desired to show both the number of cases in major groups as well as the subgroups simultaneously, we use the component bar diagram.

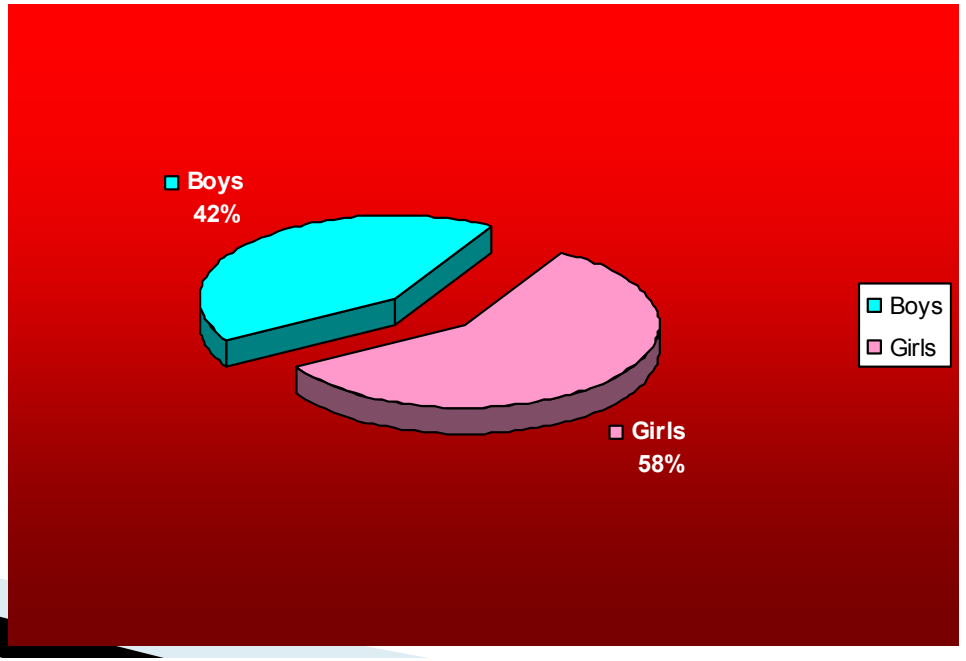
Mon	Male	female	children
JAN	299	256	143
FEB	139	278	321
MARCH	145	132	231



2) PIE DIAGRAM:

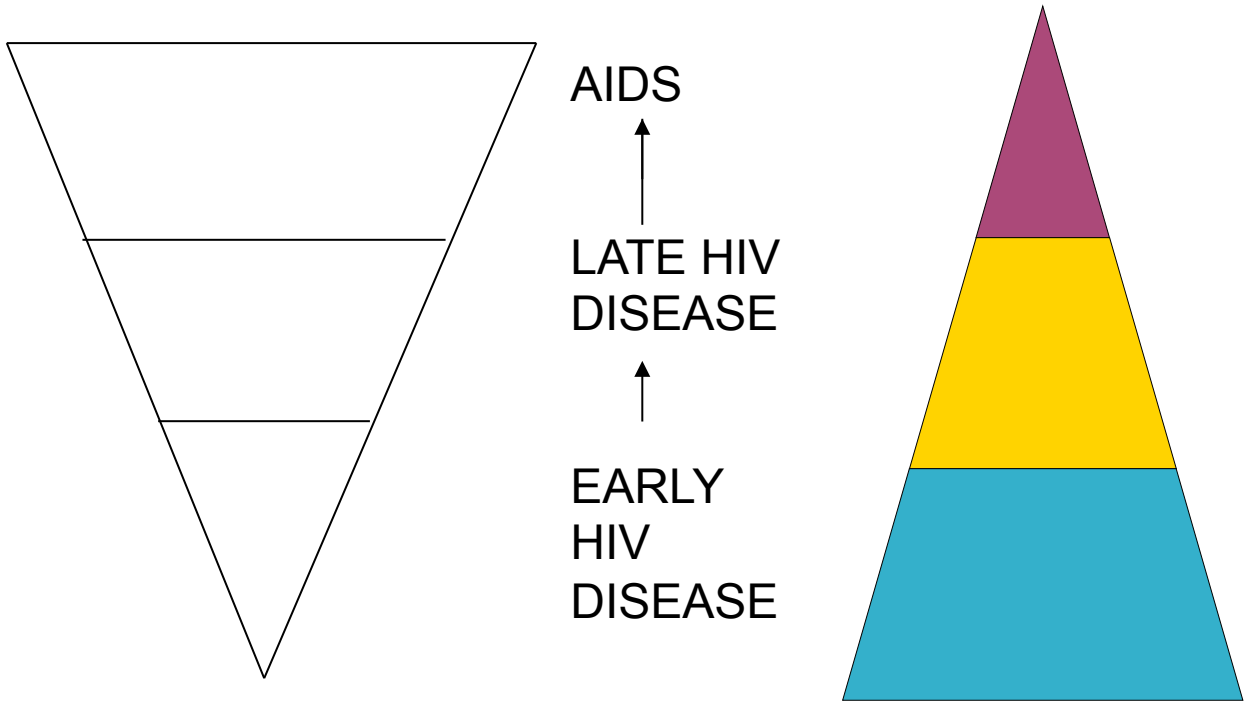
- Used to show percentage breakdowns for qualitative data.
- The entire graph looks like a pie.
- The circle is divided into different sectors corresponding to the frequencies of the variables in the distribution.
- Observation of caries in school children during camp.

Std.5	Boys	67
	Girls	94
Std.6	boys	93
	Girls	84



3) Pictogram or Picture Diagram

Pictogram of HIV infection and AIDS

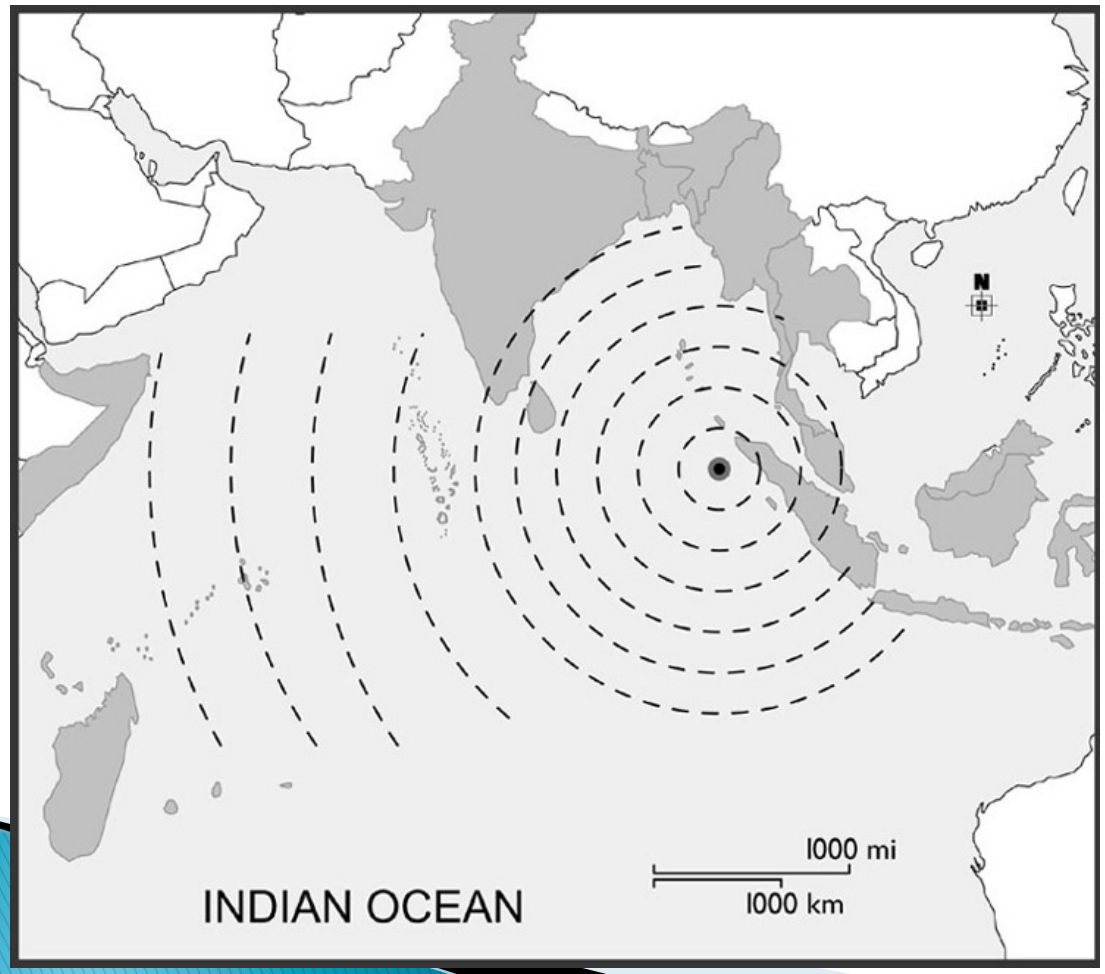


Industrially Developed countries

Developing countries

4) SPOT MAP OR CARTOGRAMS

- Used to show geographical distribution of frequencies of a characteristic



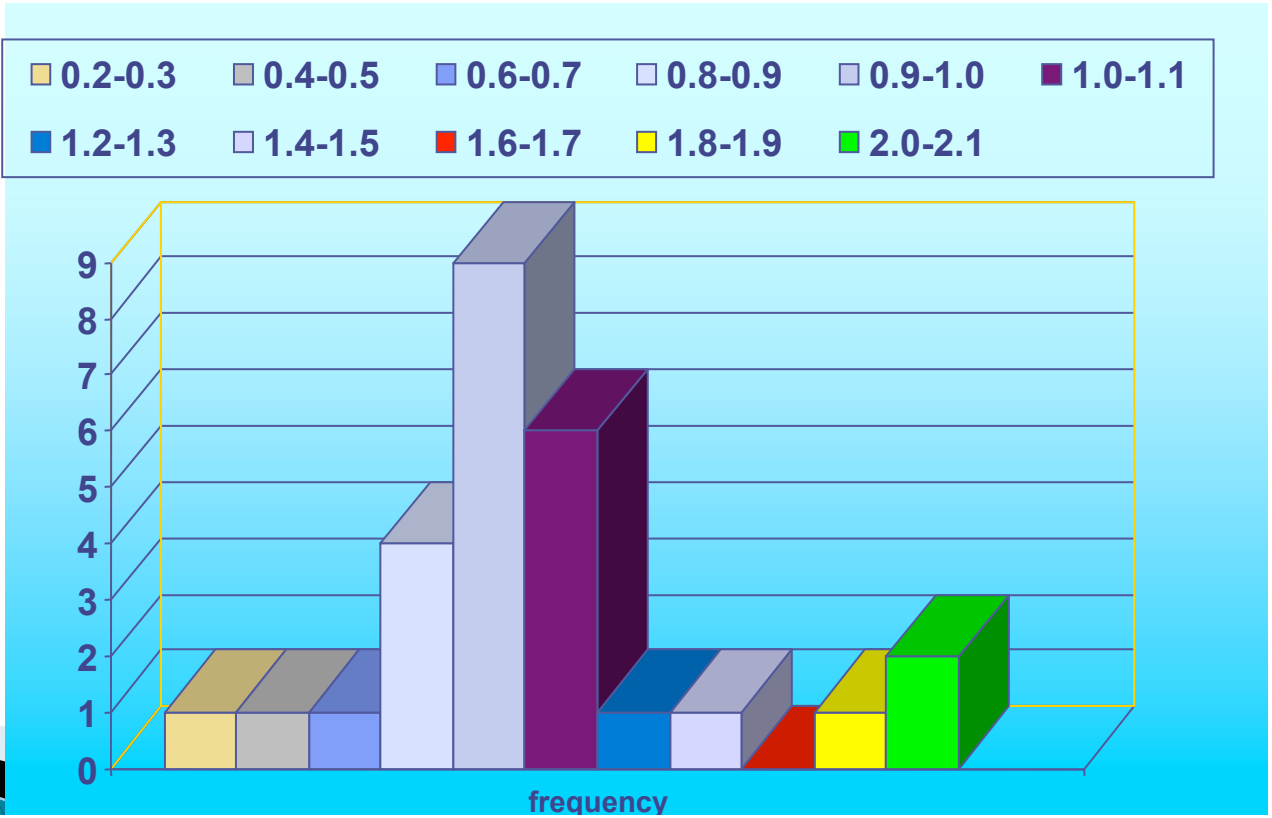
Tsunami map in December 2004

GRAPHS
FOR
QUANTITATIVE DATA

1) HISTOGRAM

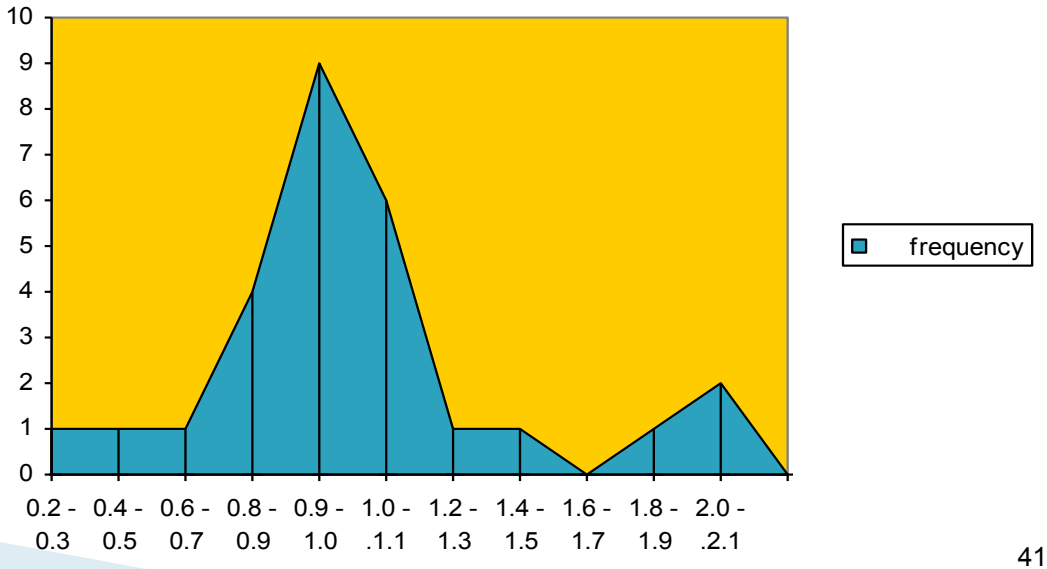
- Used to represent quantitative data of continuous type.
- It is a bar diagram without gap between bars.
- It represent a frequency distribution of fluoride concentration in ppm water supplies of 25 communities.

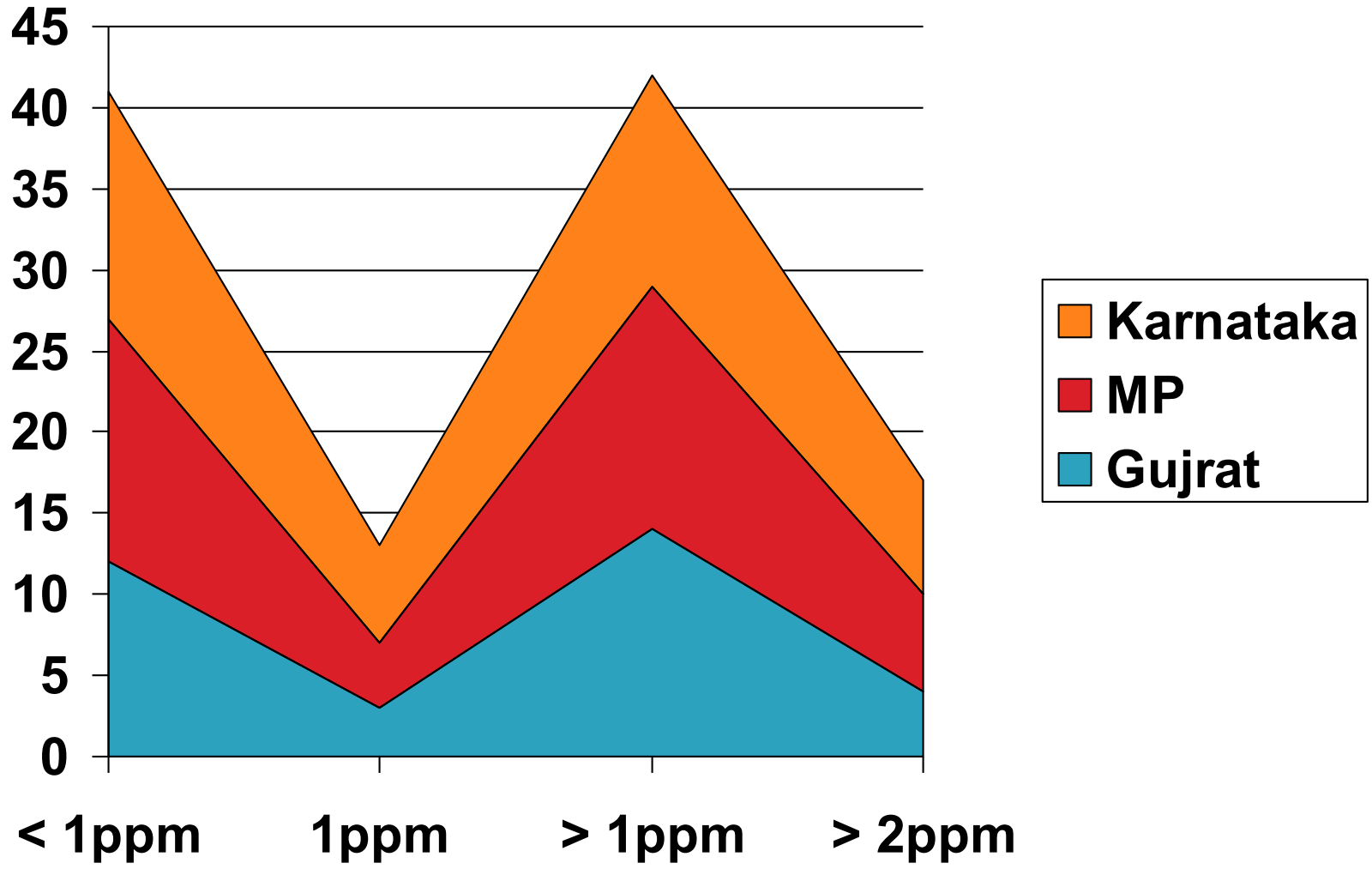
class interval	frequency
0.2 –0.3	1
0.4 – 0.5	1
0.6 – 0.7	1
0.8 – 0.9	4
0.9 –1.0	9
1.0 –.1.1	6
1.2 – 1.3	1
1.4 – 1.5	1
1.6 –1.7	0
1.8 – 1.9	1
2.0 –.2.1	2
Total	



2) FREQUENCY POLYGON

- Used to compare two or more frequency distribution for quantitative data
- To draw this, a point is marked over the mid point of the class interval, corresponding to the frequency.
- Then, straight lines connect these points.
- To compare two or more frequency distributions, lines of different types are drawn on the same graph.

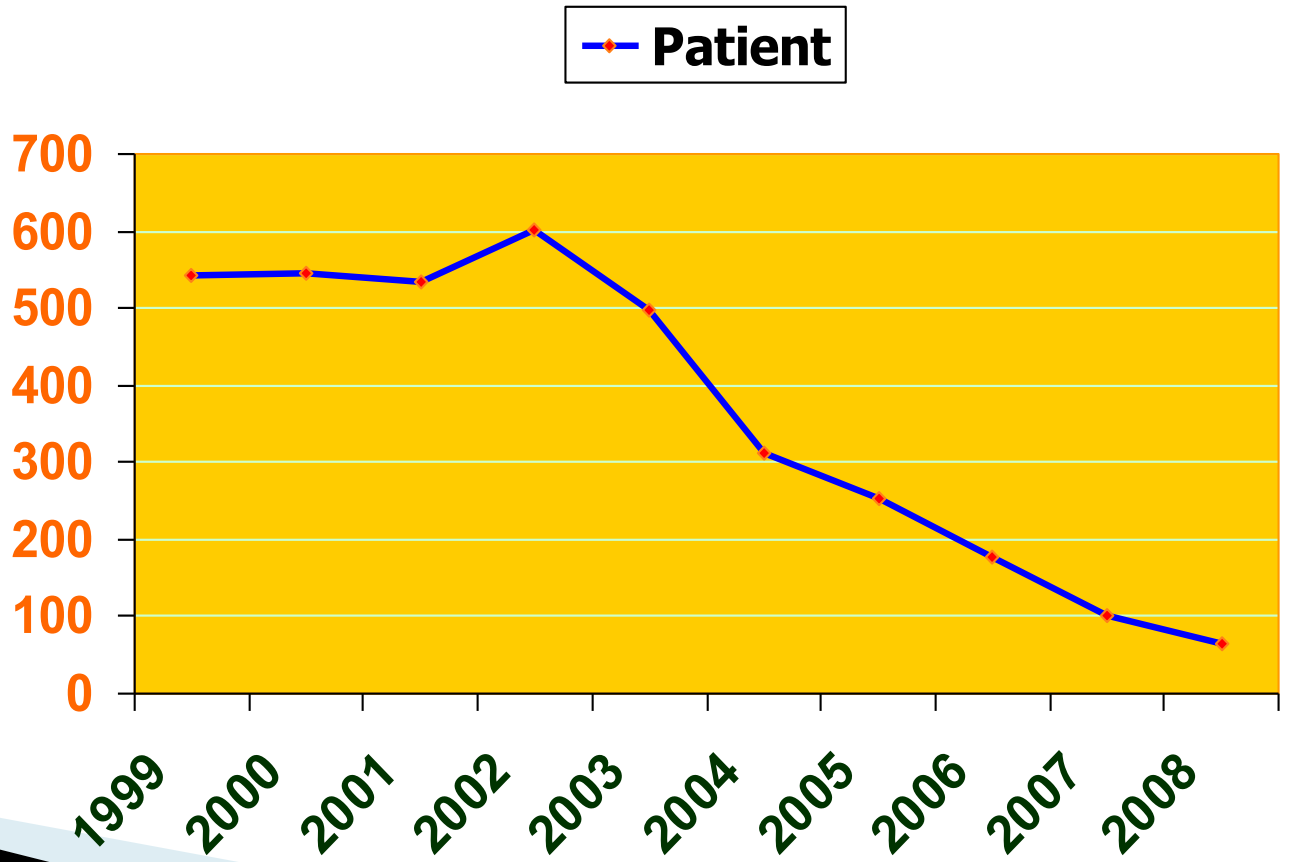



45
40
35
30
25
20
15
10
5
0
< 1ppm
1ppm
> 1ppm
> 2ppm
Karnataka
MP
Gujrat

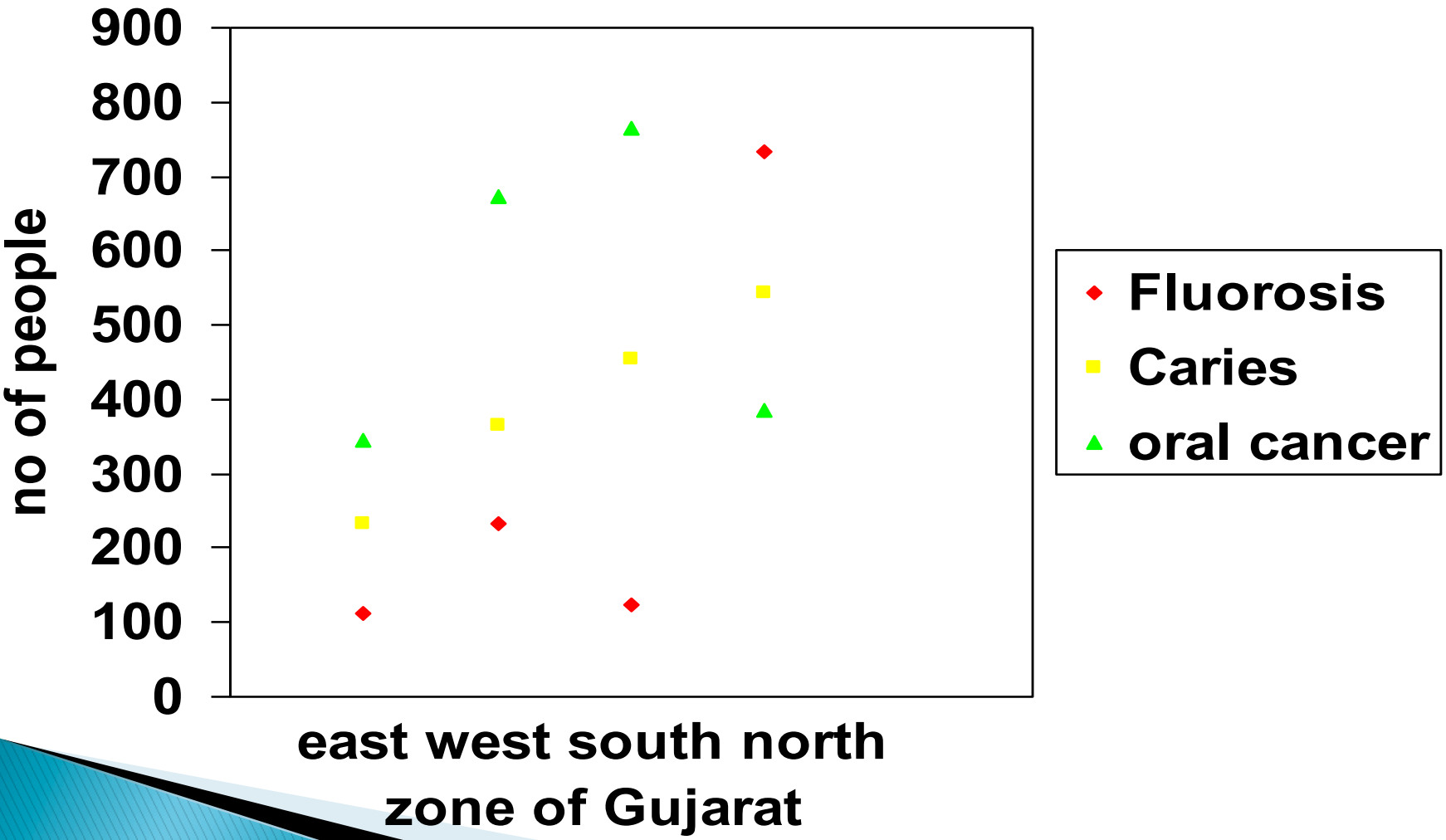
4) LINE GRAPH:

- Simplest type of diagram
- Useful to study changes of values in the variable over time.
- Number of patient at the OPD of dental clinic for 10 years.

Years	Patients
1999	543
2000	544
2001	534
2002	602
2003	498
2004	312
2005	254
2006	178
2007	102
2008	64



6) Scatter or dot diagram



MEASURES OF CENTRAL TENDENCY

- Used for comparison of two or more data series.
- For overall comparison of the distributions, the entire mass of data may be summarized using a single value.
- This single estimate of a series of data that summarizes the data is known as the parameter and one such parameter is the measure of central tendency.
- Objectives:-

  # To condense the entire mass of data

  # To facilitate comparison.

- Properties:-
  - It should be easy to understand and compute
  - It should be based on each and every item in the series.
  - It should not be affected by extreme observations.
  - It should be capable of further statistical computations.
  - It should have sampling stability.

Measures of central tendency used in dental sciences

1) Arithmetic mean 2) Median 3) Mode

1) Arithmetic mean:-

- It is the simplest measure of central tendency.
- Mean is calculated as follows;

$$\text{Mean} = \frac{\text{Sum of all the observations of the data}}{\text{Number of observations in the data}}$$

OR

$$\bar{x} = \frac{\Sigma x_i}{n}$$

where, sigma(Σ) means the sum of , x_i is the value of each observation in the data and n is the number of observations in the data.

- When the observations are small in size, simply add them up and divide by the number of observations.
- The number of decayed teeth in a group of 8 children aged 6 years are as follows: 2,4,6,3,3,4,3,7
- So, the mean number of decayed teeth for this group is calculated as

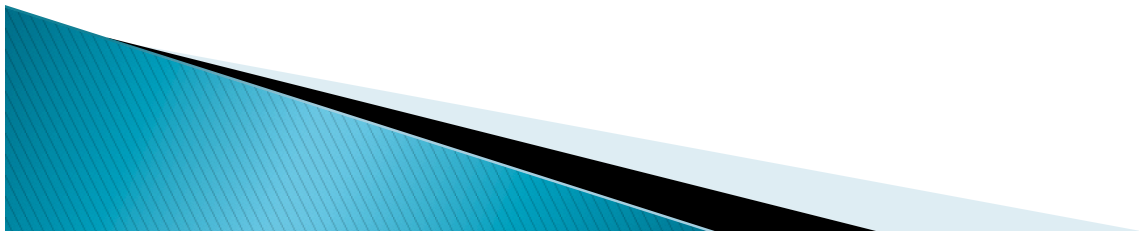
$$\text{Mean } X = \frac{2+4+6+3+3+4+3+7}{8} = \frac{32}{8} = 4 \text{ teeth}$$

MERITS OF MEAN:

- Rigidly defined
- Easy to understand and calculate
- Based upon all observation

DEMERITS OF MEAN :

- Affected by extreme values
- If one observation is missed, mean can't be calculated
- Can't calculated by inspection.



MEDIAN

- The median, by definition, is the middle value in a distribution such that one half of the units in the distribution have a value smaller than or equal to the median and one half has a value higher than or equal to the median.
- All observations are arranged in the order of their magnitude and then the middle value of the observations is selected as the median.
- When the number of observations is odd the (n +1) / 2 the value will correspond to a single value.
- When the number of observations is even, the mean of the two middle values may be taken as the median.

MERITS :

- Easy to understand and calculate
- Not affected by extreme values
- Can be located graphically

DEMERITS :

- Not based upon all the observations
- Affected much by fluctuations of sampling

- Observation 3,7,8,1,9
- Observation 4,2,7,5,8,9

- Fluctuation 2,5,21,40,60

MODE

- It is that value in a series of observations that occurs with the greatest frequency.
- When mode is ill defined, it can be calculated using the relation.
- Mode = 3 median – 2 mean.

MERITS :

- Easy to understand and calculate
- Calculated by graphically also
- Not affected by fluctuations of sampling
- Calculated both from quantitative and qualitative data

DEMERITS :

- Not based on all observations
- Some cases, mode is ill defined

- 5,2,9,14,2,5,4,9,3,5,12,10,5……..
- 6,7,8,7,6,8,6,7,8,6,8……FORMULA

MEASURES OF DISPERSION :

- Measures of central tendency give us a single value that represents the entire data.
- But, this does not adequately describe the data.
- Dispersion is the degree of spread or variation of the variable about a central value.
- The measures of dispersion help us to study the spread of the values about the central value.

The most common measures of dispersion used in dental science are

1) Range
2) Standard deviation
3) Coefficient of variation

- Range:

- difference between the value of the smallest item and the value of the largest item.
- Gives no information about the value that lie between the extreme values.

Standard Deviation

- It is also known as root mean square deviation because it is the square root of the mean of the squared deviations from arithmetic mean.
- It is a measure of the differences of each observation from the mean of all the observations.
- A small SD means a higher degree of uniformity of the observations.
- Computation of SD.
  - Calculate the mean, X , of the series
  - Take the deviations, of the items from the mean,i.e., d= X –Xi
  - Square these deviations (d^2) and obtain the total $\sum d^2$.
  - Divide $\sum d^2$ by the total number of observations, n or n–1 if sample size is less than 30.
  - obtain the square root. $S = \sqrt{\sum d^2 / n}$

Weight of 3-4 yrs of boys

Weight-Xi	Deviation - $d=X-\bar{X}i$	d^2
12	12-12.6= -0.6	0.36
12	12-12.6= -0.6	0.36
14	14-12.6=1.4	1.96
11	11-12.6= -1.6	2.56
13	13-12.6= 0.4	0.16
10	10-12.6= -2.6	6.76
15	15-12.6= 1.4	1.96
12	12-12.6= -0.6	0.36
13	13-12.6= 0.4	0.16
14	14-12.6=1.4	1.96

Here $\sum$ Xi= 126

- **So mean**

$$\bar{X} = 126/10$$

$$= 12.6$$

and $\sum d^2=16.6$

$$= \sqrt{\sum d^2 / n}$$

$$= \sqrt{16.6/ 10}$$

$$= \sqrt{1.66}$$

$$=1.358$$

Coefficient of variation

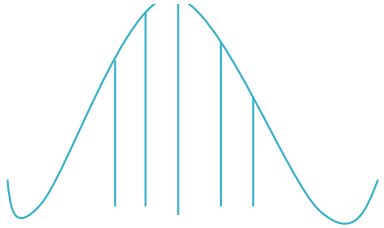
- To compare two or more series of data with either different units of measurement or either marked difference in mean, a relative measure of dispersion known as coefficient of variation is used.
- C.V. =(s x 100) / $\bar{x}$
- Higher the C.V. greater is the variation in the series of data.
- 21 years group –mean Hb 12.3654 –SD 0.9514

  6 years group–mean DMFT 1.9546 – SD 1.0562
- So C.V1 (for 21 yrs) = 0.9514 x 100 / 12.3654 = 0.07

  C.V2 (for 6 yrs) = 1.0562 x 100 / 1.9546 = 0.54

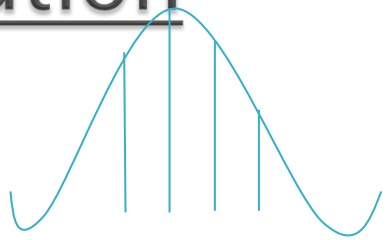
CV2 shows more variation.

THE NORMAL CURVE

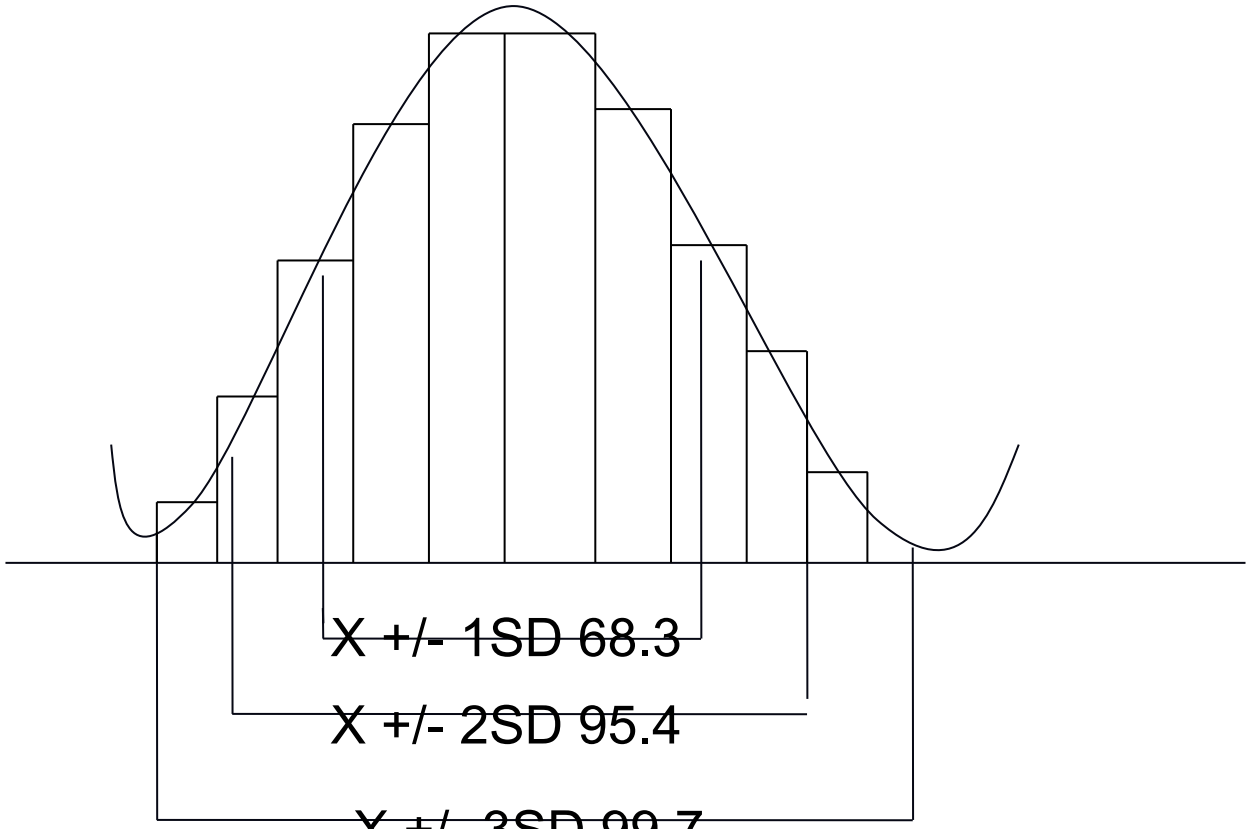


- When histogram is drawn for given frequency distribution, we generally see the height of the bars is the greatest at the middle
- Some values of the observation are above the mean and others are below it.
- If sample size is increased and the class width is narrowed and a histogram is drawn, then we see that half the observations lie above and half below the mean and all observations are distributed equally on either side of the mean.
- A distribution with this nature or shape is called normal distribution or the Gaussian distribution.
- Thus, with the help of mean and SD we can describe a frequency distribution.

Properties of the Normal Distribution

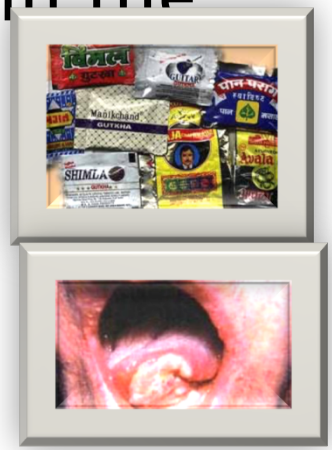


- The normal curve is bell shaped.
- The curve is symmetrical about the middle point.
- The height of the curve is maximum at the mean.
- all the three measures of central tendency, the mean, median and mode coincide.
- maximum number of observations are at the value of the variable corresponding to the mean and the numbers of observations gradually decrease in either side with few observations at the extreme points of central tendency, viz. the mean, median and mode coincide.
- The area under the curve is represented in terms of a relationship between the mean and the standard deviation which is a follows :
- Mean ± 1 S.D. covers 68.3 % of the observation(32% outside the range)
- Mean ± 2 S.D. covers 95.4 % of the observation(4.55% outside)
- Mean ± 3 S.D. covers 99.7 % of the observation (0.27% in 100)


X +/- 1SD 68.3
X +/- 2SD 95.4
X +/- 3SD 99.7

VARIABLES(V)

- A variable is a state,condition,concept or event whose value is free to vary within the population.
- Independent variables….tobacco
- Dependent variables….oral cancer
- Confounding or intervening variables….nutrition
- Background variables…age, social status



TESTS OF SIGNIFICANCE

- The methodologies of statistics that deal with techniques to know how far the differences between the estimates of different samples is due to sampling variation or otherwise is known as testing of hypothesis.
- Terms needs to understand:

1) Null hypothesis
2) Alternative hypothesis
3) Level of significance
4) Degree of freedom
5) Standard error

1. Null hypothesis: [H_o]
   - First step in testing of hypothesis is to set up an appropriate hypothesis with the problem.
   - The null hypothesis asserts that there is no real difference between the two groups under consideration and the difference found is accidental and arises out of sampling variations.

2. Alternative hypothesis: [H_1]
   - In case of rejection of the null hypothesis we need another hypothesis.
   - Usually stated that there is a difference between the two groups being compared.

- Test of significance is performed.
- Accordingly null hypo. Rejected / accepted.

Four possible ways of interpreting the result:

- The H_o is true and our test accepts it
- The H_o is false and our test rejects it
- The H_o is true still it is rejected
  [type I error]
- The H_o is false still it is accepted
  [type II error]

Inference	accept it	Reject it
Ho is true	Correct	Type I error
Ho is false	Type II error	correct

3. Level of significance

- The maximum probability of rejecting a null hypothesis is always some % [usually 5% to 1%].
- It is called level of significance and it is determine by probability level ‘p’.
- ‘p’ is range from 0 to 1.
- ‘p’ = 0 ---means no chance of an event happening/ impossible.
- ‘p’ = 1---means 100% chance of an event
- p + q = 1 (q – chances of not happening)

4. Degree of freedom :

- The degree of freedom is defined as the number of independent members in the sample.
- $df = (r-1) \times (c-1)$

5. Standard error :

- It is not a error or mistake but it is a measure of chance variation
- Take many samples of same size from the population.
- Assess the variability of such means
- Mean of these means is the population mean. This variability can be estimated from a single study.
- The difference of individual means from this population or grand mean is SE.

- Formula $s_{\bar{X}} = \sum (\bar{X} - \mu)^2 / n-1$

- X is individual mean , μ is mean of sample means

Steps involved in testing of a hypothesis

- 1) State an appropriate null hypothesis for the problems
- 2) Calculate the suitable statistics using the standard error; t, x^2, Z etc.
- 3) Determine the degrees of freedom for the static.
- 4) Find the probability level p corresponding to the test static using the relevant tables
- 5) The null hypothesis is rejected /accepted.
- To measure significance of difference different tests are 'Z' test, 't' test and chi square test – 'x^2' test

TESTS IN TEST OF SIGNIFICANCE

parametric
(normal distribution)

non parametric
(not follow normal distribution)

quantitative

1.Student ‘t’ test.

i] unpaired

ii] paired

2.Z test [for large number].

3. ANOVA

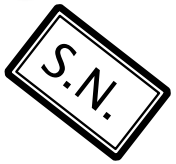
4. ANCOVA.

qualitative

1. Z proportional test

2. χ^2 test [chi]^2

1.STUDENT'S UNPAIRED t- TEST



- Applied to unpaired data of independent observations made on **two different groups from two populations**
- W.S. Gossett – 1958.
- Criteria : random samples
  Quantitative data
  Variable normally distributed
  Sample size less than 30

Application of ‘t’ test :

1. Comparison of two means of small independent sample
2. Comparison of sample mean and population mean.
3. The test statistic is given by

$$t = \frac{X1^2 - X2^2}{SE}$$

$$SE = \sqrt{1/n1 + 1/n2}$$

e.g. Upper incisor width in mm of 2 groups

Group I :
8,7.5,8.2,7.5,7.3,8.3,6.8,7.2,6.7,7.6

Group II :
8.9,5.4,6.7,8.8,6.5,5.2,8.1,7.8,9.4,9.1

Any statistical significance difference exists by chance?

- Null hypothesis : no significant diff. bet. 2 group in relation to mean upper incisor width

I	II
N1 = 10	n2 = 10
Mean X1 = 7.5100	x2 = 7.5900
S1 = 0.5425	S2 = 1.5429

$$SE = \sqrt{1/n1 + 1/n2}$$

$$= 0.5369$$

- $t = \frac{X1^2 - X2^2}{SE}$

  $t = 7.51 - .7.59 / 0.5369 = \underline{0.1490}$

  Now df = $n_1 + n_2 - 2 = 18$

  In table 't' at 5% value is $\underline{2.1010}$

  t cal < t tab

  Not significant

  Null hypothesis is accepted.

2.PAIRED t TEST

- For the same group for two times
- Patient is examine before and after a treatment
- Two measurements are available for each patient
- Called the paired setup
- To know significant difference bet paired set of measurements, the test is called paired t test.

Application

- Applied to paired data of independent observations, from on sample only when each individual gives pair of observation
- The test statistic is given by
- $t = \bar{x} / SE_{\_}$

$\bar{X}$ = mean of the différences

9 patients having BP drug given and then BP measured again

No	Before	After	Diff.
Mean	125	122	3
SD	0.1560	1.254	1.09
SE			0.58

$t = 3 / 0.58 = 5.17$

At 5% LOS, the table value with n-1 = 8

$t = 2.31$

t cal [5.17] > t tab [2.31]

Highly significant

Hence null hypothesis is rejected

So injection lower the BP at 5% LOS

3.Student's T or Z test

- Ratio between observed difference of 2 means and SE of means
- WS Gosset in 1958
- Criteria
  - Quantitative data
  - Normally distributed
  - Random samples
  - Sample more than 30

The test statistic is given by Z or T =

$$\frac{\text{observation - mean}}{\text{SD}} = \frac{x-\bar{x}}{\text{SD}}$$

- Summary values for caries experience.

Group	n	Mean DMFT	SD
Control	32	1.1578	1.2
Experimental	32	2.9645	2.2

Find whether difference between mean DMFT is significant

5.Chi Square Test … χ^2

- It is an alternative method of testing the significant different between two or more proportions
- Application

1. Used to find significant difference in two or more than two proportions.
2. Used as test of association between two events in binomial or multinomial samples

Smoking and oral cancer

School levels and oral health

Knowledge of breast cancer and opinion about mammography

Criteria

- Random samples
- Qualitative data
- Lowest observation frequency is not less than 5

$$\chi^2 = \Sigma\ (O - E)^2\ /\ E$$

O = Observed frequency

E = Expected frequency

Steps

- State null and alternative hypothesis
- Make a contingency table
- Determine expected frequency (E)
- E = (row total X column total) /sample total
- Find O – E

$\chi^2 = \Sigma$ (O – E)^2 / E for each cell

- To find the efficacy of drug from the data given below :

Group	Died	Survived	Total
Control	10	25	35
(on placebo)	5.25	29.75	
Experiment	5	60	65
On drug	9.75	55.25	
Total	15	85	100

E = (row total X column total) /sample total

1. Expected number (E) of the died in control group = 15/100 x 35 = 5.25

   χ^2 value of this cell = (O – E)^2 / E

   = (10–5.25)^2 /5.25

   = 4.2978.

2. Expected number (E) of the survived in control group=85/100 x 35 = 29.75

   χ^2 value of this cell= (25 – 29.75)^2 / 29.75

   = 0.7584

3. χ^2 value for the died in experiment group= 2.3140 X^2 value
4. χ^2 value for the survived in experiment group = 0.4083

Total χ^2 value = 4.2976
+0.7584+2.3140+0.4083
= 7.7783

On referring table of χ^2 , as 1 degree of freedom, the value of χ^2 under probability 0.05 is 3.841 and under o.o2 is 5.412.

Calculated value χ^2 is higher the table value so it is highly significant.

Null hypothesis is rejected.

Drug is effective.

- Calculate degree of freedom

i.e. df = (C –1) (r –1)

C = column and r = rows

Sum of all X^2 values of all cells

$$= \Sigma (O - E)^2 / E$$

Refer X^2 with any LOS and compare with calculated value of χ^2

Draw a conclusion

6.ANOVA–Analysis of variance

- Three or more groups
- Experimental situations
  - Type of treatment
  - Dose of drug
  - Types of material used

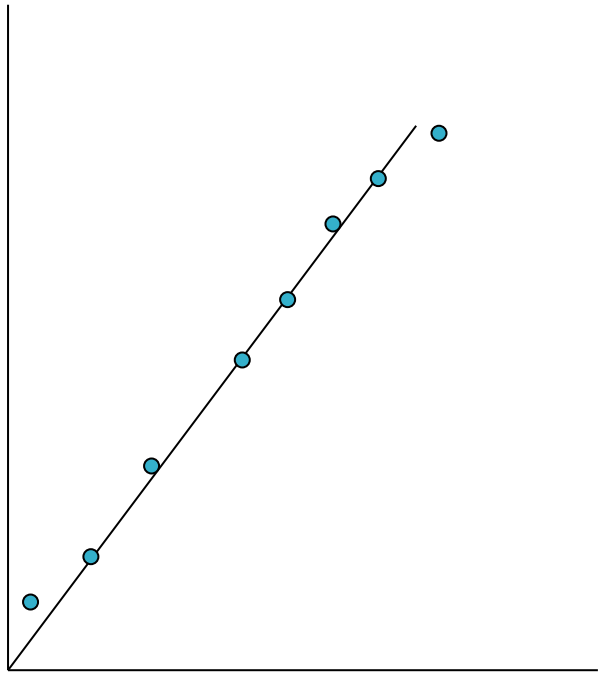
CORRELATION & REGRESSION

Correlation – r

- Correlation is the relationship between two sets of variables.
- Degree of relationship between two variables --- correlation coefficient– represented by "r".
- 'r' is range from -1 to +1, $-1<r<+1$.
- Type and extend of relationship between the two variable can be obtained by scatter diagram.

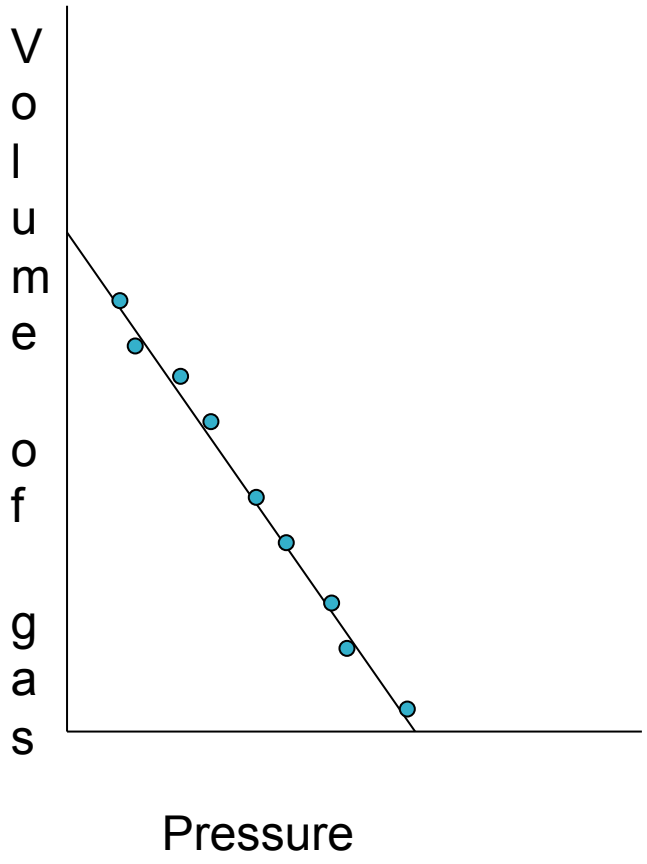
TYPES OF CORRELATION

1. <u>Perfect positive correlation</u>
   - correlation coefficient (r) = +1
   - both variable rise & fall in the same proportion
   - height and weight

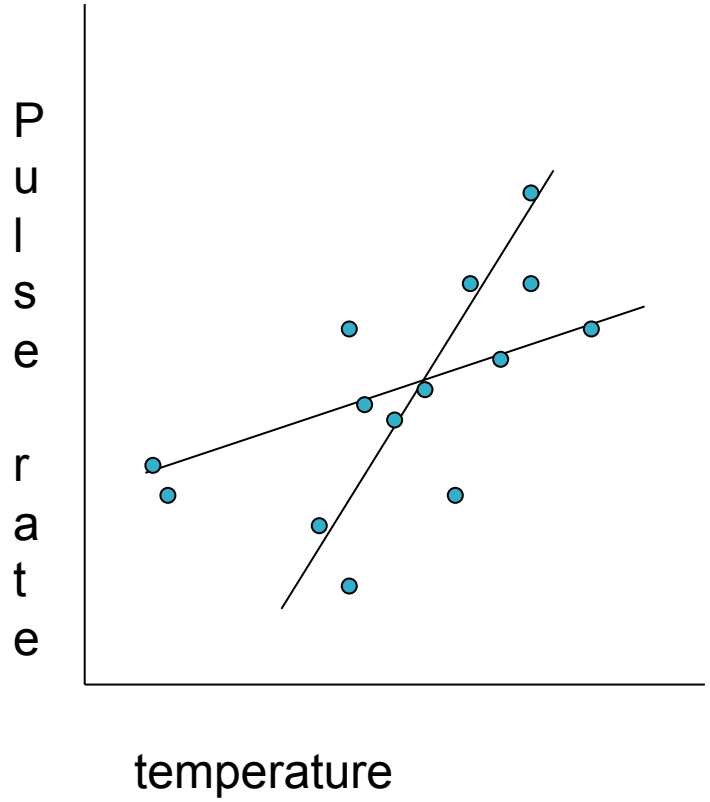


2. Perfect negative correlation

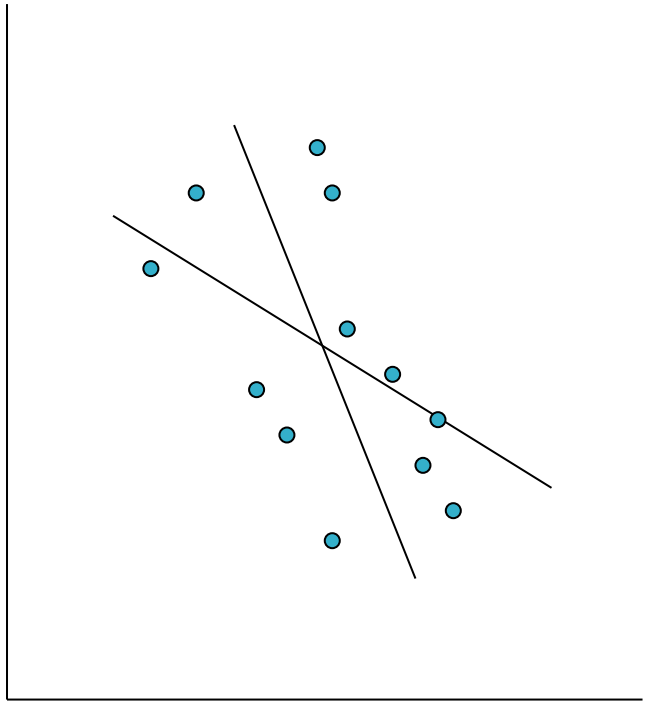
- Correlation coefficient (r) = -1
- When one variable rises , the other falls
- Pressure and volume of gas



3. Moderately positive correlation
   - (r) lie between 0 to +1
   - scatter will be there around an imaginary mean line rising from lower extreme values of both variables

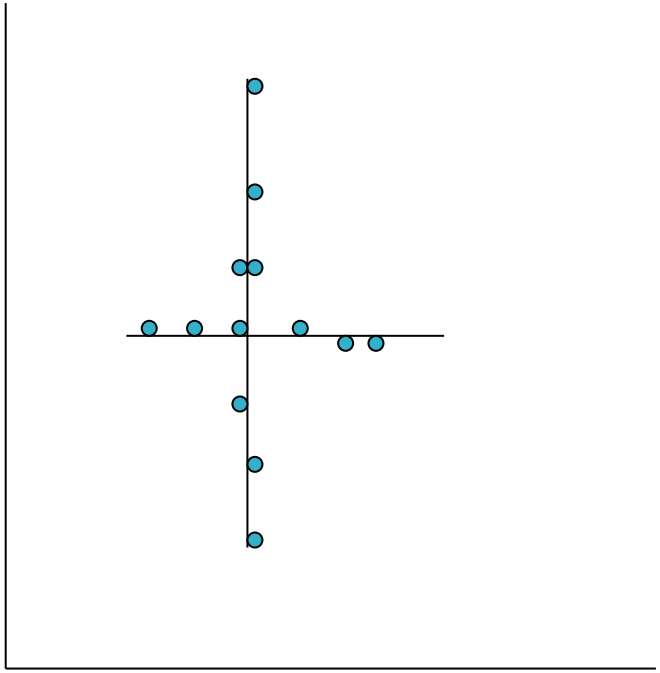


4. Moderately negative correlation
   - (r) lie between −1 and 0
   - scatter will be there around an imaginary mean line rising from upper extreme values of both variables



5. Absolutely no correlation

- Value of correlation coefficient is zero
- No relationship bet the two variables



REGRESSION – b

- Correlation(r) only measures the degree of relationship bet X & Y variables- not give an idea about the changes in which variable leads changes in another
- This is done in REGRESSION
- Variable : causes change- *independent*

changes in manner with a
changes in independent variable-
dependent

- Regression is the change of the dependent variable with respect to change in the independent variable.
- Regression is measured by regression coefficient- represented by 'b'
- Regression coefficient gives the amt of increase or decrease, in dependent variable with the change in the independent variable
- Correlation coefficient..
- $r = \sum (X-\bar{X})(Y-\bar{Y}) / \sqrt{\sum X-\bar{X})^2 (Y-\bar{Y})^2}$

- Regression means going backwards
- Regression is a statistical methods with the help of which we can estimate unknown values of one variable from the known values of another variable, provided the variables are correlated.
- Regression provides estimate of values f the dependent variable from the value of the independent variables
- Regression analysis is widely used in almost all scientific disciplines, natural, physics and social sciences

- Correlation gives- degree and direction of relationship bet the two variable
- Regression analysis- value of one variable on the basis of the other variable

- **BIOSTATISTICS**
- Define biostatistics. Write in detail the uses of biostatics in public health in detail.____(7)
- Presentation of statistical data.____(7)
- Variable_____(5)
- Simple random sampling._____(7)
- Define sampling. Classify sampling. Enumerate any one of the sampling.____(7)
- Types of samples____(5)
- Random sampling._____(5)
- 't' test / chi square test—(5)
- Pie diagram._____(5)
- Standard deviation.____(5)
- Sources of data._____(5)
- Histogram____(5)
- Mean, median, mode.____(5)
- Median_____(5)
- Types of diagrams____(5)
- Bar chart____(5)
- Correlation and regression____(5)


thank u