

Basics in Biostatistics



Dr. Jyothi Conjeevaram
Professor
Dept. of Community Medicine
NMC, Nellore

WHAT IS STATISTICS ?

Statistics is the science of collecting, organizing, presenting, & analyzing numerical data for the purpose of assisting in making more effective decisions

WHAT IS BIO-STATISTICS?

Biostatistics is the application of statistics to the biologic sciences, medicine and public health.

The **tools of statistics are employed in** many fields:

business, education, psychology, agriculture, economics, ... etc.

We use the term **biostatistics** to distinguish this particular application of statistical tools and concepts when the data analyzed are derived from the biological science and medicine.

DATA:

- ▶ **The raw material of statistics is data.**
- ▶ We may define data as figures. Figures result from the process of counting or from taking a measurement.
- ▶ *For example:*
 - ▶ - When a hospital administrator counts the number of patients (counting).
 - ▶ - When a nurse weighs a patient (measurement)

Sources of Data:

We search for suitable data to serve as the raw material for our investigation.

Such data are available from one or more of the following sources:

1- Routinely kept records.

For example:

- Hospital medical records contain immense amounts of information on patients.

2- External sources.

The data needed to answer a question may already exist in the form of:

published reports,

commercially available data banks, or

the research literature, i.e. someone else has already asked the same question.

3- Surveys:

The source may be a survey, if the data needed is about answering certain questions.

For example:

If the administrator of a clinic wishes to obtain information regarding the mode of transportation used by patients to visit the clinic, then a survey may be conducted among patients to obtain this information.

4- Experiments


Frequently the data needed to answer a question are available only as the result of an experiment.

For example:

If a nurse wishes to know which of several strategies is best for maximizing patient compliance,

she might conduct an experiment in which the different strategies of motivating compliance are tried with different patients.

APPLICATION & USES OF DATA

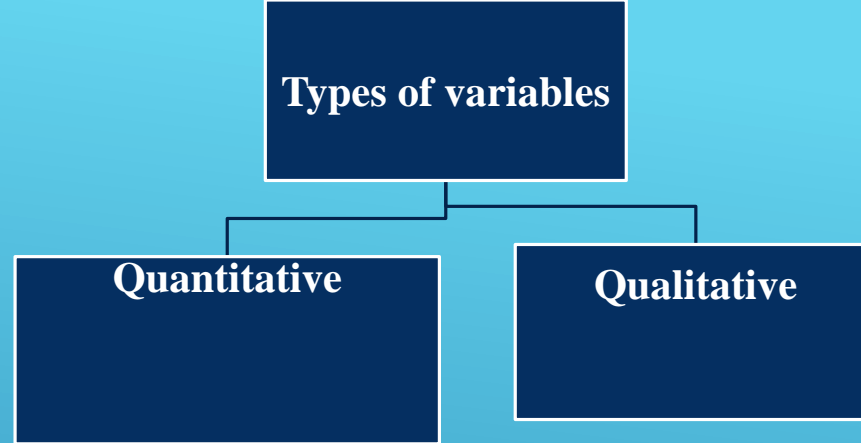
- ▶ In physiology & anatomy : to define normal limits
 - ▶ In pharmacology: to compare the action of two drugs. To find the efficacy of new drug with standard drug
 - ▶ In medicine: to find association between two attributes- cigarette smoking & lung cancer
 - ▶ In community medicine- evaluation of any health programmes
- 

A VARIABLE:

It is a characteristic that takes on different values in different persons, places, or things.

For example:

- heart rate,
- the heights of adult males,
- the weights of preschool children,
- the ages of patients seen in a dental clinic.



Quantitative Variables

It can be measured in the usual sense.

For example:

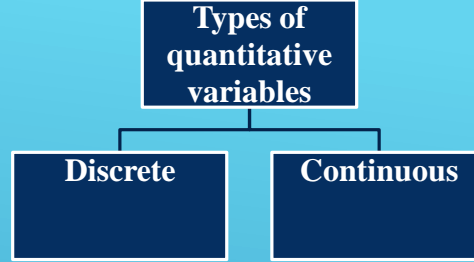
- the heights of adult males,
- the weights of preschool children,
- the ages of patients seen in a dental clinic.

Qualitative Variables

Many characteristics are not capable of being measured. Some of them can be ordered or ranked.

For example:

- classification of people into socio-economic groups,
- Color of eyes: blue, green, black, brown etc.
- Exam results: pass or fail



A discrete variable

is characterized by gaps or interruptions in the values that it can assume.

For example:

- The number of daily admissions to a general hospital,
- The number of decayed, missing or filled teeth per child in an elementary school.

A continuous variable

can assume any value within a specified relevant interval of values assumed by the variable.

For example:

- Height,
- weight,
- skull circumference.

No matter how close together the observed heights of two people, we can find another person whose height falls somewhere in between.

Type of Measurements

There are four types of measurements of data. They are :

Nominal Data : “Nominal” scales could simply be called “labels.” All of these scales are mutually exclusive (no overlap) and none of them have any numerical significance.

Characterized by data that consist of names, labels, or categories only. The data cannot be arranged in an ordering scheme (such as low to high) **Example: Gender: 1=Male, 2=Female**

Ordinal Data :

involves data that may be arranged in some order, but differences between data values either cannot be determined or are meaning less. It is the order of the values is what’s important and significant, but the differences between each one is not really known. For example, is the difference between “OK” and “Unhappy” the same as the difference between “Very Happy” and “Happy?” We can’t say. Ordinal scales are typically measures of non-numeric concepts like satisfaction, happiness, discomfort, etc. “

Eg: Rating of the Cancer (Stage I, II, III etc.): Very Unhappy, Unhappy OK, Happy, Very happy

Type of Measurements

Interval level data : scales are numeric scales in which we know not only the order, but also the exact differences between the values. **The classic example of an interval scale is Celsius temperature** because the difference between each value is the same. For example, the difference between 60 and 50 degrees is a measurable 10 degrees, as is the difference between 80 and 70 degrees.

Time is another good example of an interval scale in which the increments are known, consistent, and measurable.

. “Interval” itself means “space in between,” which is the important thing to remember—interval scales not only tell us about order, but also about the value between each item.

The problem with interval scales: they don't have a “true zero.” For example, there is no such thing as “no temperature.” Without a true zero, it is impossible to compute ratios.

Ratio level Data :

The interval level modified to include the natural zero starting point.

For values at this level, differences and ratios are meaningful. They tell us about the order, they tell us the exact value between units, AND they also have an absolute zero—which allows for a wide range of both descriptive and inferential statistics to be applied.

Everything above about interval data applies to ratio scales + ratio scales have a clear definition of zero.

Good examples of ratio variables include height and weight. Ratio scales provide a wealth of possibilities when it comes to statistical analysis. These variables can be

Example: AGE, Weight, Height, HB

How to present a data

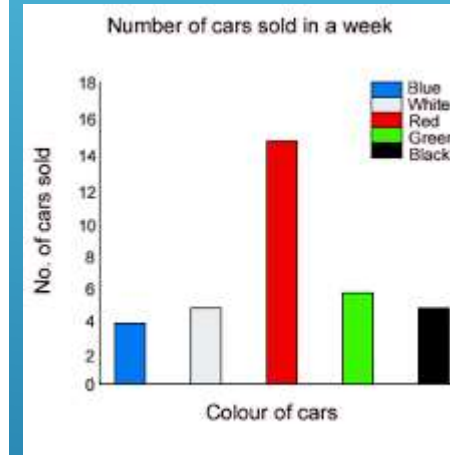
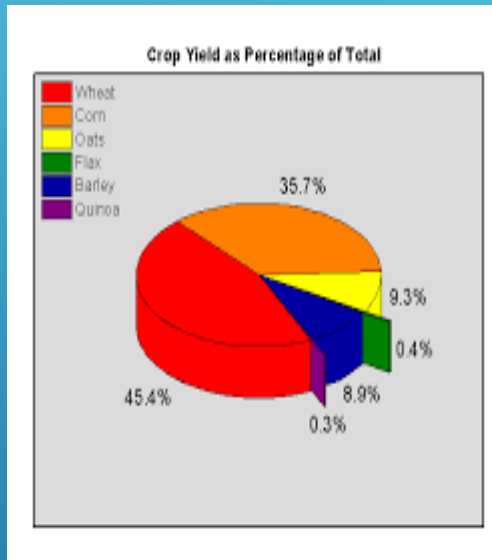
- ▶ Simple tables
- ▶ Frequency distribution tables
- ▶ Histogram
- ▶ Bar diagrams
- ▶ Pie diagram
- ▶ Line diagram
- ▶ Pictograms
- ▶ Shaded maps

State	Total Population
Uttar Pradesh	199,581,477
Maharashtra	112,372,972
Bihar	103,804,637
West Bengal	91,347,736
Andhra Pradesh	84,665,533
Madhya Pradesh	72,597,565

Scores
40, 43, 54, 62, 88, 31, 94, 83, 81, 75, 62, 53, 62, 83, 90, 67, 58, 100, 74, 59

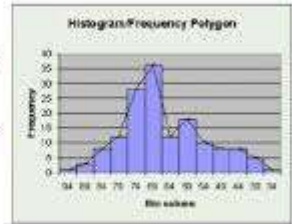
Numerical Scores	Letter Grade
>=80	A
70-79	A-
60-69	B
50-59	C
40-49	D
<40	F

Number	No. of Students	Cumulative Frequency
>=80	7	7
70-79	2	9
60-69	4	13
50-59	4	17
40-49	2	19
<40	1	20



Activity #2
Frequency Polygon

Plot frequencies versus variable values and then join the points with straight lines



Use the data from activity #1 to construct a frequency polygon.



TABULATION OF DATA

- ▶ Tabulation is the 1st step before the data is used for analysis /interpretation.
- ▶ Tables can be simple or complex
- ▶ General principles to be borne in mind before constructing a table are:
 - tables should be numbered
 - brief self explanatory titles
 - clear & concise headings for rows & columns
 - data must be presented according to size/importance: chronologically
alphabetically
geographically
 - percentages, if needed should be placed as close as possible
 - table should not be too large
 - vertical arrangement is preferred to horizontal
 - footnotes

SIMPLE TABLES

State	Total Population
Uttar Pradesh	199,581,477
Maharashtra	112,372,972
Bihar	103,804,637
West Bengal	91,347,736
Andhra Pradesh	84,665,533
Madhya Pradesh	72,597,565

FREQUENCY DISTRIBUTION TABLE

- ▶ In a frequency distribution table, the data is first split into class intervals and the frequency which occurs in each group is shown in a different column


Scores 40, 43, 54, 62, 88, 31, 94, 83, 81, 75, 62, 53, 62, 83, 90, 67, 58, 100, 74, 59

Grading Policy	
Numerical Scores	Letter Grade
>=80	A
70-79	A-
60-69	B
50-59	C
40-49	D
<40	F

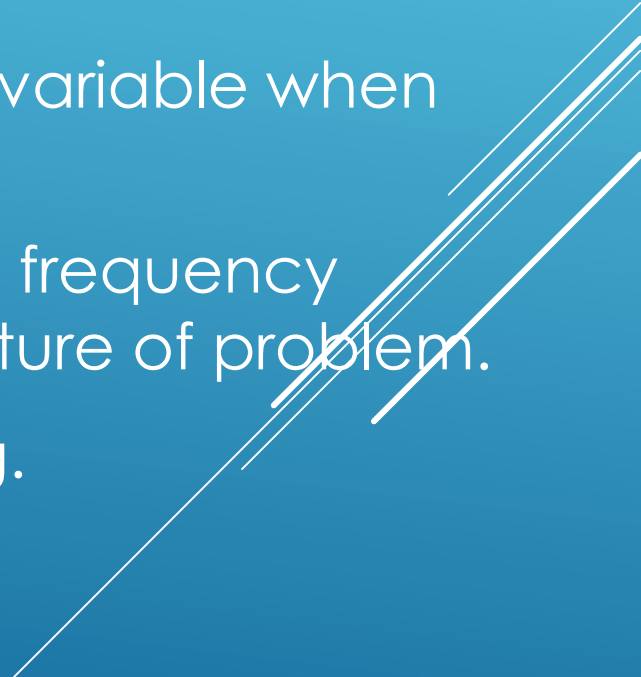
Frequency Distribution Table		
Number	No. of Students	Cumulative Frequency
>=80	7	7
70-79	2	9
60-69	4	13
50-59	4	17
40-49	2	19
<40	1	20

Intervals	Tally Marks	Frequency
10-19		2
20-29		2
30-39		6
40-49		8
50-59		9
60-69		8
70-79		1
80-89		2
90-99		1

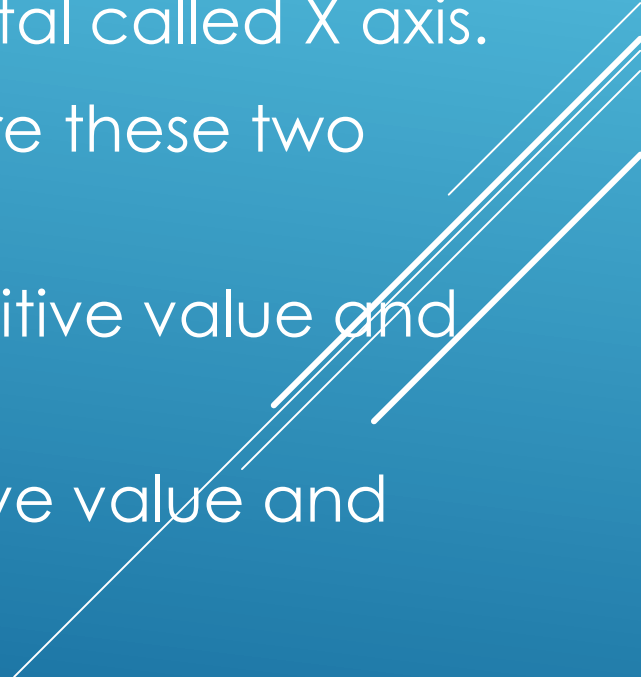
GRAPHIC PRESENTATION OF DATA

- ▶ A Graphical representation is a visual display of data and statistical results using plots and charts.
 - ▶ It is more often and effective than presenting data in tabular form.
 - ▶ There are different types of graphical representation and which is used depends on the nature of the data and the nature of the statistical results.
 - ▶ It is used in many academic and professional disciplines but most widely so in the field of mathematics, medicine and the science.
 - ▶ Graphical representation helps to quantify, sort and present data in a method that is understandable to a large variety of audience.
- 


GRAPHIC PRESENTATION OF DATA

- ▶ Graphs enable us in studying the cause and effect relationship between two variables.
 - ▶ Graphs help to measure the extent of change in one variable when another variable changes by a certain amount.
 - ▶ Graphs also enable us in studying both time series and frequency distribution as they give clear account and precise picture of problem.
 - ▶ Graphs are also easy to understand and eye catching.
- 

GENERAL PRINCIPLES OF GRAPHIC PRESENTATION

- ▶ There are some algebraic principles which apply to all types of graphic representation of data.
 - ▶ In a graph there are two lines called coordinate axes.
 - ▶ One is vertical known as Y axis and the other is horizontal called X axis.
 - ▶ These two lines are perpendicular to each other. Where these two lines intersect each other is called '0' or the Origin.
 - ▶ On the X axis the distances right to the origin have positive value and distances left to the origin have negative value.
 - ▶ On the Y axis distances above the origin have a positive value and below the origin have a negative value.
- 

TYPES OF DIAGRAMS

- ▶ 1. BAR DIAGRAMS
 - ▶ 2. LINE DIAGRAMS
 - ▶ 3. HISTOGRAM
 - ▶ 4. PIE DIAGRAM
- 
- A decorative graphic consisting of several parallel white lines of varying lengths and positions, arranged in a diagonal pattern from the bottom right towards the top right of the slide.

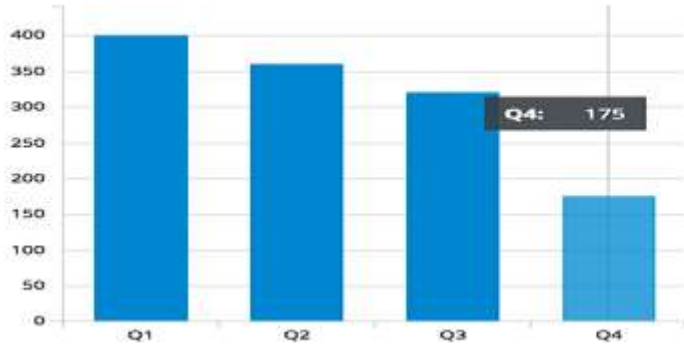
BAR DIAGRAM

- ▶ Popular media of presenting statistical data because they are very easy to prepare and enable values to be compared visually.
- ▶ Used for qualitative type of data
- ▶ A graph showing the differences in frequencies or percentages among the categories of a nominal or an ordinal variable.
- ▶ The categories are displayed as rectangles of equal width with their height proportional to the frequency or percentage of the category.
- ▶ Length of the bar is proportional to the magnitude to be represented
- ▶ A bar graph will have two axes.
- ▶ One axis will describe the types of categories being compared and the other will have numerical values that represent the values of the data.
- ▶ The bars can be plotted vertically or horizontally.

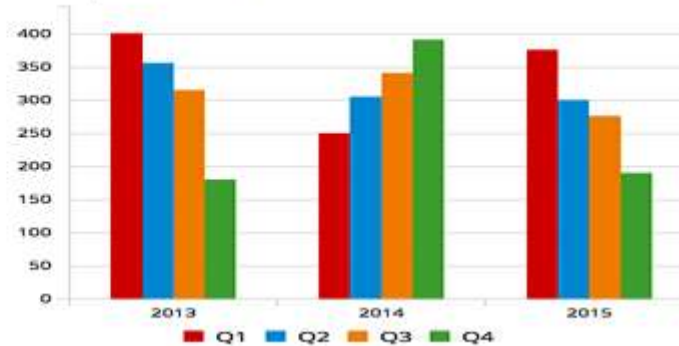
TYPES OF BAR CHARTS

- ▶ Simple bar charts: can be vertical or horizontal
- ▶ Multiple bar charts
- ▶ Component /stacked bar chart

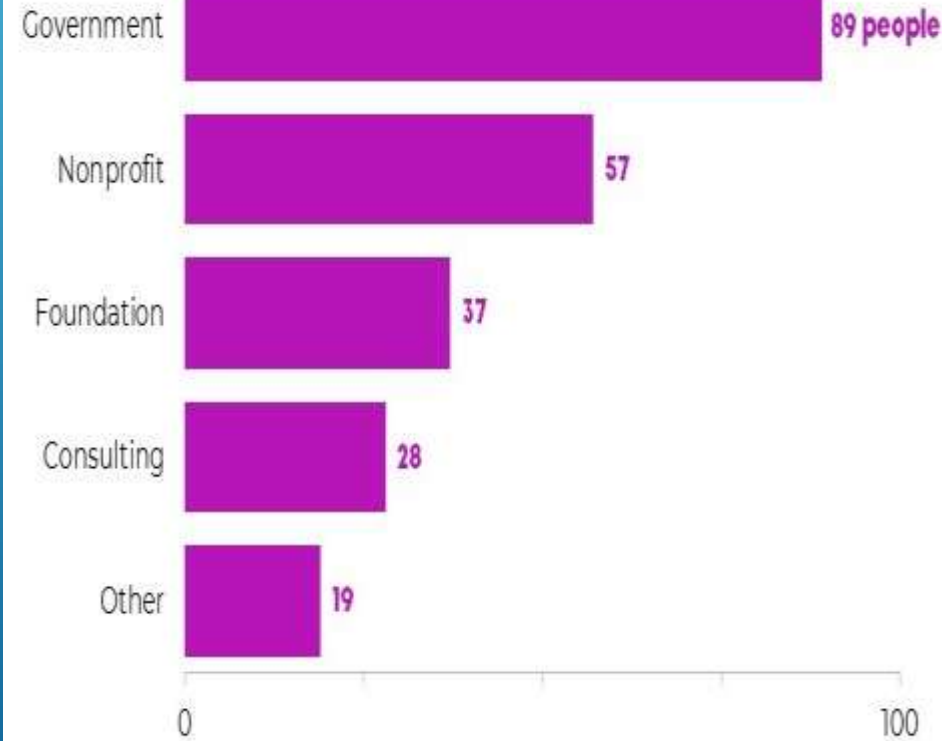
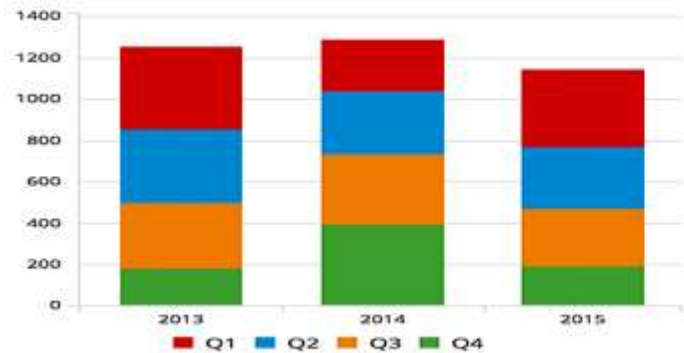
Vertical Bar Chart



Grouped Vertical Bar Chart

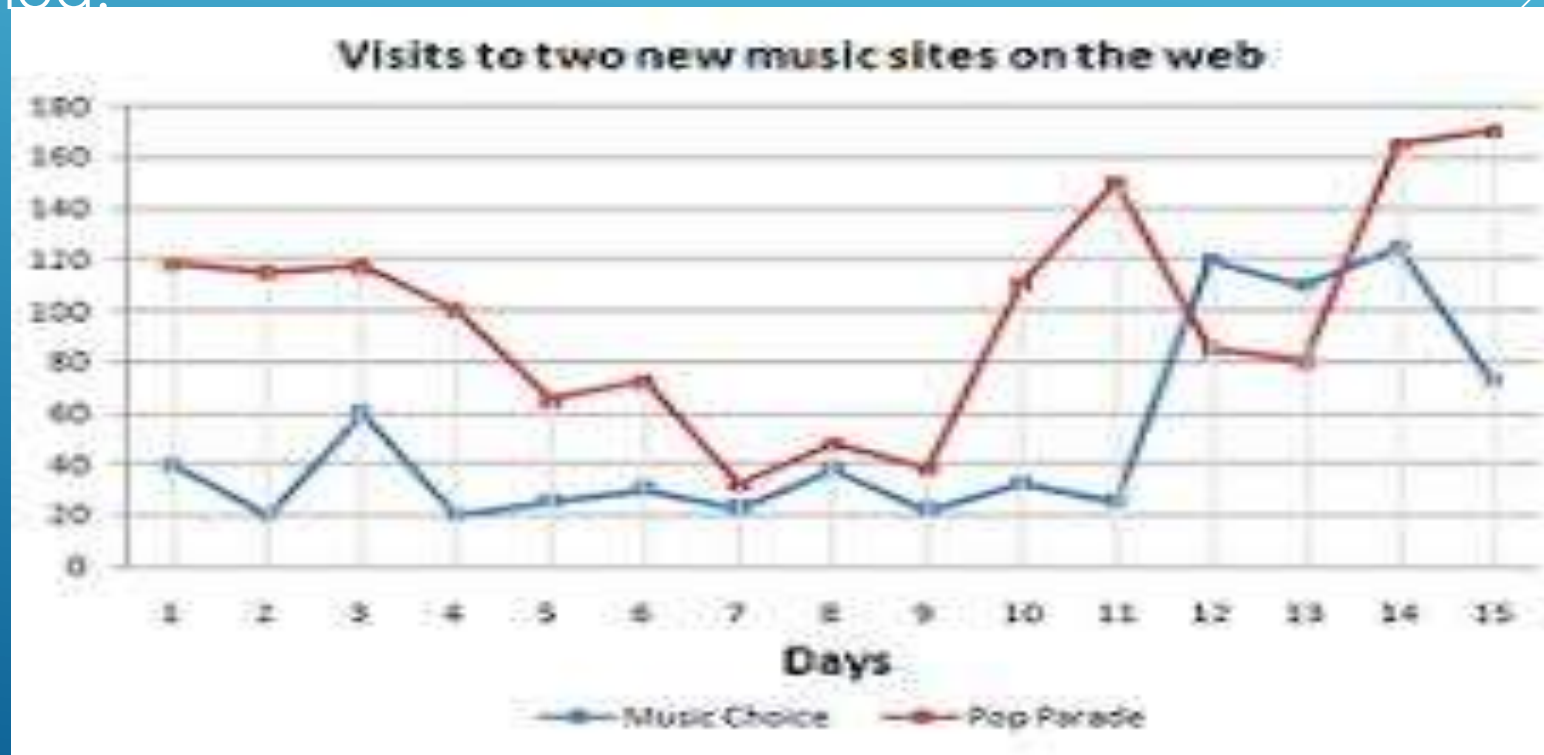


Stacked Vertical Bar Chart



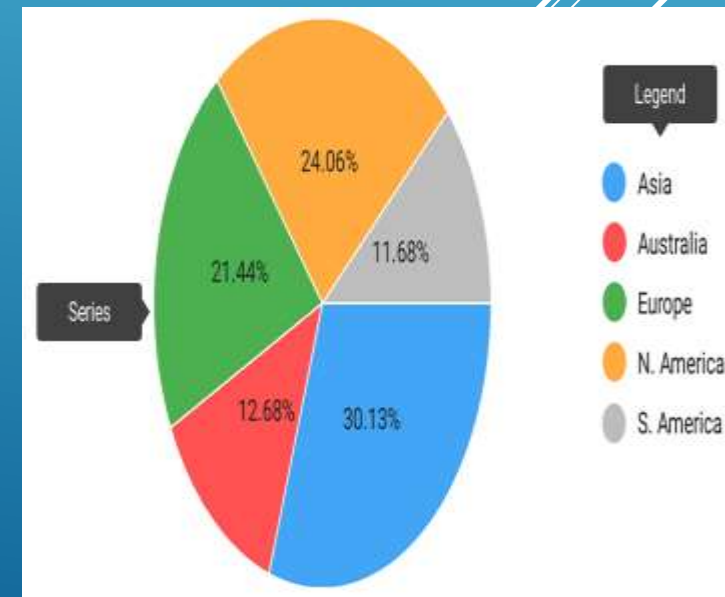
LINE DIAGRAMS

- ▶ Used for quantitative data
- ▶ Show the trend of events with the passage of time
- ▶ Helps in assessing the trends and displaying data on epidemics/outbreaks in the form of epidemic curves.
- ▶ Advantage of line diagram is comparisons can be made at a given period of time or over the time period.



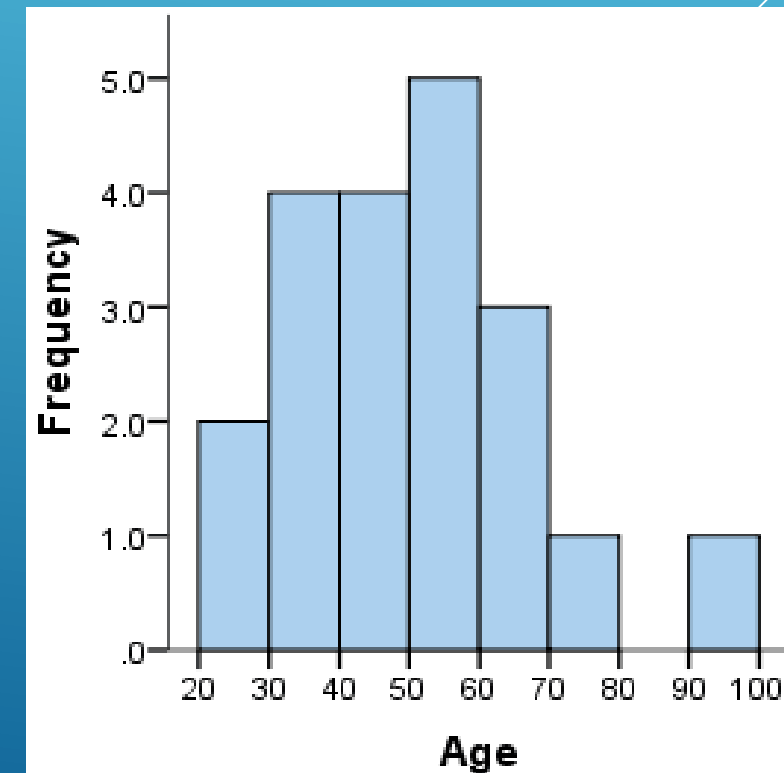
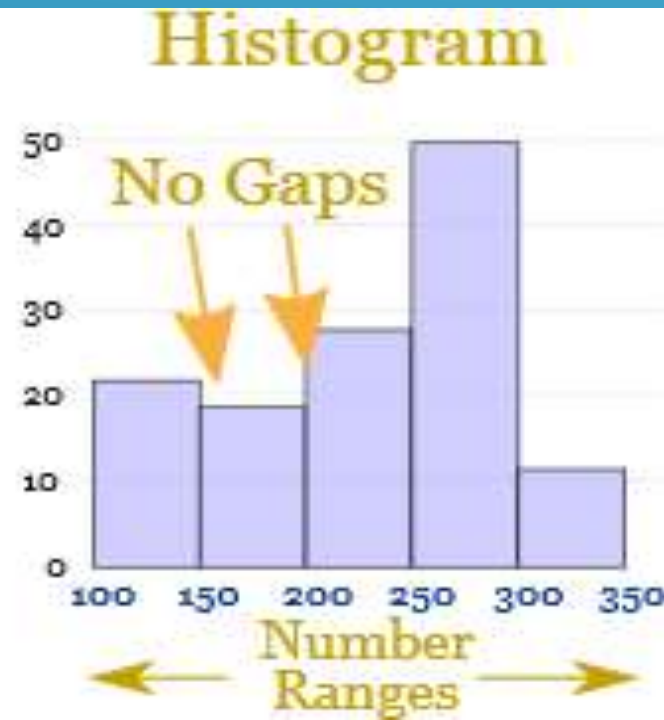
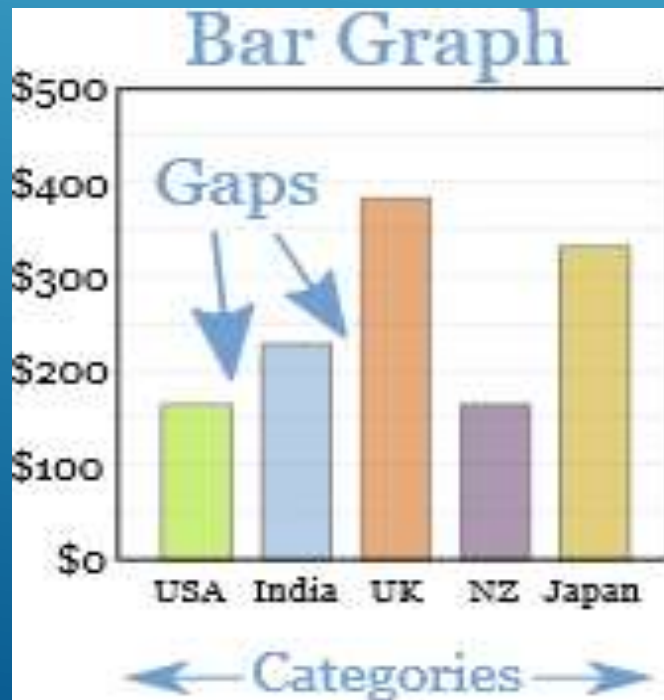
PIE CHARTS

- ▶ Used to display qualitative data
- ▶ Instead of bars we use segments which are constructed after calculating the angles
- ▶ Can be used when the total categories are b/w 2 and 6
- ▶ Total percentage should add up to 100
- ▶ Different colors for diff. segments/diff. shades of same color



HISTOGRAM

- ▶ Used for quantitative continuous data
- ▶ We plot the class intervals on x-axis and frequencies on the y-axis
- ▶ The area of each block is proportional to the frequency
- ▶ Differs from bar diagram in that the blocks are continuous without gaps b/w them



OTHERS

- ▶ Pictograms are popular method of presenting data to the man in the street
- ▶ Small pictures/symbols are used to present data



- ▶ Shaded maps

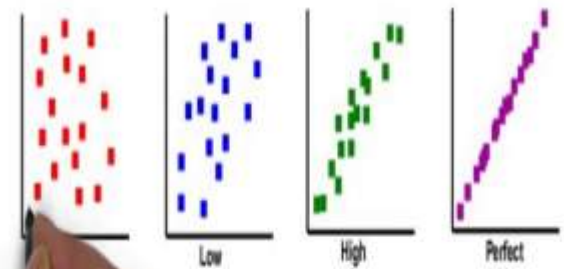


SCATTER DIAGRAM

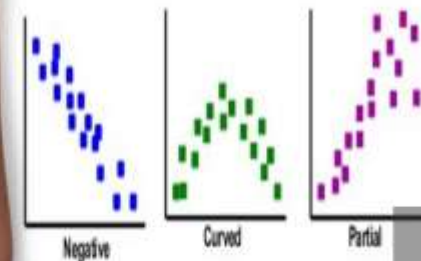
- ▶ Gives quick visual display of association b/w 2 variables
- ▶ Both are on continuous/discrete scale
- ▶ Dependent variable on Y-axis and
- ▶ Independent variable on X-axis

Scatter Diagram - How do I use it? - Correlation

Degrees of correlation:



on:



EXERCISES

I. **SIMPLE BAR DIAGRAM**

illustrate the following data in vertical bar diagram

a. Average births/woman in 1990

kerala: 2

india : 4

b. IMR in 1992 (per 1000 live births)

kerala : 18

india:80

c. Distribution of blood groups of patients with HTN

Blood groups	frequency	percentage
A	232	42.8
B	201	37.1
AB	76	14
O	33	6.1

HORIZONTAL BARS

Illustrate the following data in horizontal bar diagram

1. Fall in reported poliomyelitis cases after pulse polio immunization in the years 1994, 1995, 1996

585 cases in 1994

415 in 1995 &

272 in 1996



MULTIPLE BAR DIAGRAM

- ▶ Distribution of malnourished status in males & female children in an Anganwadi center:

	malnourished	normal
Male	6	44
Female	11	39

- ▶ Age-specific mortality rates of India in 1981 & 1991

age in yrs	1981	1991
15-19	90.4	76.1
20-24	246.9	234
25-29	232.1	191.3
30-34	167.7	117
35-39	102.5	66.8
40-44	44	30.6

PIE CHARTS

- ▶ Distribution of patients according to blood groups

Blood groups	percentage
A	42.8
B	37.1
AB	14
O	6.1

- ▶ Obstetric causes of maternal mortality in India:
 - haemorrhage – 38%
 - sepsis – 11%
 - abortions – 8%
 - obstructed labour – 5%
 - HTN disorders -5%
 - other conditions – 34%



LINE DIAGRAM

- ▶ Illustrate the progress in MTP in India by line chart using the following data:

year	number in thousands
1972-73	20.1
73-74	41.2
74-75	105.5
75-76	214.2
76-77	278.8
77-78	247
78-79	317.7
79-80	360.8
80-81	388.4
81-82	426.5





Thank you

ABIGAIL