



Collecting Data



Presented By:
Dr.S.Singh

Grammatical point

Datum -- singular

Data -- plural

DATA

- A dictionary defines data as facts or figures from which conclusions may be drawn.
- A collective recording of observations either numerical or otherwise is called data

Statistics is a tool for converting *data* into



- But where then does *data* come from ?
- How is it gathered ?
- How do we ensure its accurate ?
- Is the data reliable?
- Is it representative of the population from which it was drawn?

Methods of Collecting Data

- **Direct Observation**
- **Experiments, and**
- **Surveys.**

Methods

Mailing paper questionnaires to respondents, who fill them out and mail them back

Having interviewers call to respondents on the *telephone* and ask them the question in a telephone interview

- Sending the interviewers to the respondent's home or office to administer the questions in *face-to-face* interviews

Depending upon the nature of the variable data is classified into 2 broad categories

- **Qualitative Data**
- **Quantitative Data:-**
 - 1. discrete**
 - 2. continuous**

- **Qualitative data :-**

(characterized by words)

when the data is collected on the basis of attributes or qualities like sex, malocclusions, cavity etc.

- **Quantitative data** :- (*characterized by numbers*)

when the data is collected through measurement, like arch length, fluoride concentration etc.

- **Discrete data** :-

when the variable under observation takes only fixed values like whole numbers.

- **Continuous data** :-

if the variable can take any value in a given range, decimal or fractional.



5 (Quantity) Happy (Quality) Kids

Quantitative: numbers breadth generalizability

Qualitative: words depth specific

Remember, "**Not everything that counts can be counted.**"

Quantitative methods – Qualitative methods

Quantitative	Qualitative
Surveys Questionnaires	Focus groups
Tests	Unstructured interviews
Existing databases	Unstructured observations

Common data collection methods

- Survey
- Case study
- Interview
- Observation
- Group assessment
- Expert or peer reviews
- Portfolio reviews
- Testimonials
- Tests
- Photographs, videotapes, slides
- Diaries, journals, logs
- Document review and analysis

Are the data reliable and valid?

- **Validity:**

Are you measuring what you think you are measuring?

- **Reliability:**

if something was measured again using the same instrument, would it produce the same (or nearly the same) results?

- The main sources of data are :-
 - 1) surveys
 - 2) experiments
 - 3) records in OPD
- Data can be collected through either :-
 - 1) primary source
 - 2) secondary source

- **Primary Source :-**

Here the data is obtained by the investigator himself. This is a first hand information.

- ***Advantages :-***

Precise information and reliable.

- ***Disadvantages :-***

Time consuming, expensive.

- **Primary data can be obtained using :-**

- 1) Direct personal

- 2) Oral examination

- 3) Questionnaire method

- **Secondary Source :-**

The data already recorded is utilized to serve the purpose of the objectives of study.

Ex:- The records of opd of the clinics.

Data Gathering Techniques

- **Observation**
- **Questioning**

Data domains

Cognitive -- paper and pencil

Affective -- interview

Psychomotor -- observation

Data Presentation

The objective of classification of data is to make the data simple, concise, meaningful and interesting and helpful in further analysis.

There are 2 methods of presenting data:-

1) Tabulation

2) Diagrams

- **Tabulation :-**

Is the first step before the data is used for analysis or interpretation.

- In the process of tabulation the following type of classification are encountered.

- 1) Geographical i.e area wise
- 2) Chronological i.e on the basis of time
- 3) Qualitative i.e. according to attribute
- 4) Quantitative i.e. in terms of magnitude

TABLE FORMAT :

Number of Primary Health Centres in India at 5-year Intervals during 1970-1990

YEAR	No.of PHCs
1970	5015
1975	5293
1980	5484
1985	7284
1990	18,981

- **Classification by Space :-**

1) Data are classified by location of occurrence

2) Arrangement of set of categories in alphabetical order of the terms defining these categories,

3) In the order of their geographical location may be found to be suitable in many cases.

Table 2

Number of Literates in Rural and Urban Areas in India (state wise list).

STATES	No.of literates (in thousands)	
	RURAL	URBAN
Andhra Pradesh	14,821	10,119
Assam	7921	1709
Bihar	20,358	6485
Maharashtra	22,164	20,774
Gujarat	12,096	9179
Kerala	16,443	6228
Madhya Pradesh	14,464	9027
Tamil Nadu	17,424	12,958
Karnataka	12,267	8812
Orissa	10,303	2608
Punjab	6253	3699
Rajasthan	8189	5428
Uttar Pradesh	33,079	13,968
West Bengal	20,337	12,382
TOTAL	216,119	123,376



- **Chronological** i.e. On the basis of time :-

1) In this case data are classified by time of occurrence of the observations

2) Arrangement of categories is almost always in chronological order

Table 3

Number of Primary Health Centres in India at 5-year Intervals during 1970-1990

YEAR	No.of PHCs
1970	5015
1975	5293
1980	5484
1985	7284
1990	18,981

- Classification by **attribute** :-
 - 1) When the data represent observations made on a qualitative characteristic the classification in such a case is made according to these qualities.
 - 2) Alphabetical arrangement of categories may be suitable for general purpose table
 - 3) In the case of special purpose table arrangement may be made in the order of importance of these categories

Table 4

Patients who attended Leprosy Clinics during 1993

Type of leprosy	No. of patients
Tuberculoid	604
Lepromatous	272
Indeterminate	72
Borderline	48
TOTAL	996

Classification by the **size of observations** :-

1. when the data represent observations of some characteristic on a numerical scale, classification is made on the basis of the individual observations.

2. The range of observations is suitable divided into smaller divisions called class intervals.

3. The numerical scale adopted may be either discrete or continuous.

Table 5

Distribution of Single Live Born with respect to Birth Weight

Birth Weight (grams)	No.of children (n)
< 1500	24
1500 - 1999	60
2000 - 2499	422
2500 and above	3202
TOTAL	3708

Principles (in tabulation of data) :-

1. Every table should contain a title, should be concise and meaningful.
2. The tables should be numbered
3. The heading of columns or rows should be clear and concise .eg: ht in cm, age in years , wt in kg etc
4. The number of class intervals should be sufficient to condense the data bringing out their significant features .
5. The class intervals should be at equal width

6. Uniform size class intervals are preferable
7. Sometimes open end class intervals are used
8. The class intervals should be clearly defined to avoid ambiguity .eg- 0-4, 5-9, 10-14 etc.
9. Units of measurements should be specified.
10. If the data is not original , the source of the data should be mentioned at the bottom of the table.
11. Groups should be tabulated in ascending or descending order.

12. If certain data is omitted or excluded deliberately, the reason for same should be given.
13. For many arithmetic calculations the midpoint of each class interval will be used as a representative of each item in that interval.
14. The interval should be so chosen that the midpoint of each interval is approximately the average of the items in that interval.
15. It will be convenient if mid points and limits are whole numbers.

16. The class intervals should be same throughout table except in case of age

- 0-<1 (infant up to 1yr)
- 1-4 (toddlers 1to 5yrs but not completed 5 yrs)
- 5-14 (School children < 15yrs)

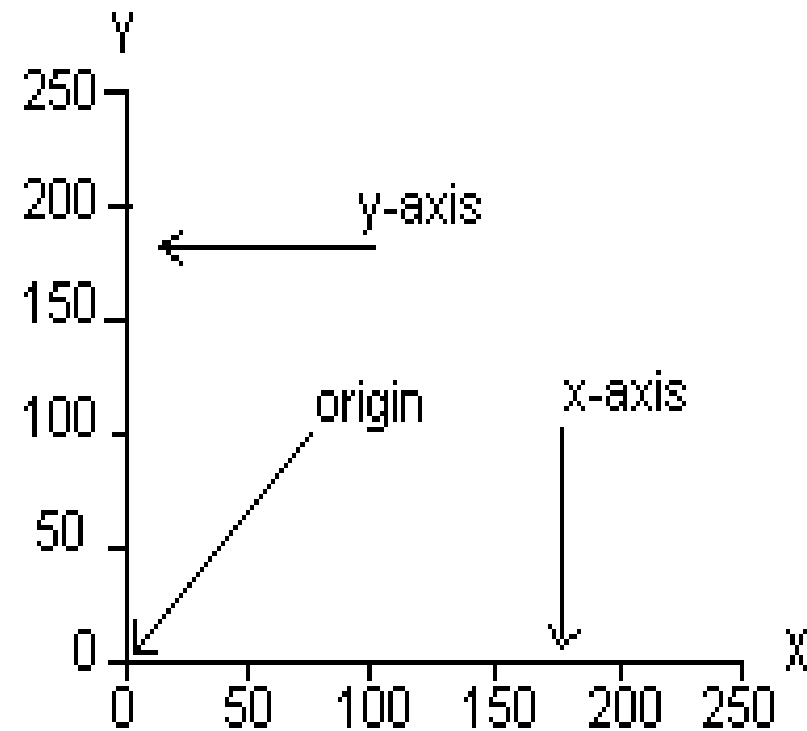
Presentation by **Graphs and **Diagrams**:-**

Diagrams and graphs are extremely useful
because

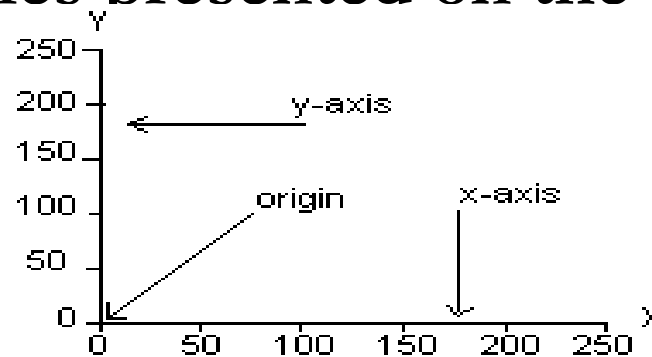
- 1. They are attractive to the eyes**
- 2. Give a birds eye view of the entire data**
- 3. Have a lasting impression on the mind of the layman**
- 4. Facilitate comparison of data.**

Basic rules in the construction of diagrams and graphs :-

1. Every diagram must be given a title
2. It should be simple
3. The vertical axis is always labeled as the 'y' axis. It is also "ordinate".
4. The horizontal axis is always labeled as 'x' axis . It is also called "abscissa".
5. The x axis and y axis meet at right angles at a point called origin (o)
6. The values of variables are presented on the x axis and the frequency on y axis



7. The number of lines drawn in any graph should not be many so that the diagram does not look clumpy.
8. The scale of presentation for the x axis and y axis should be mentioned at the right hand corner of the graph
9. The scale of division of two axis should be proportional and the division should be mark along with the details of the variables and frequencies presented on the axis.



Presentation of quantitative data is through graphs, the common graphs in use are:-

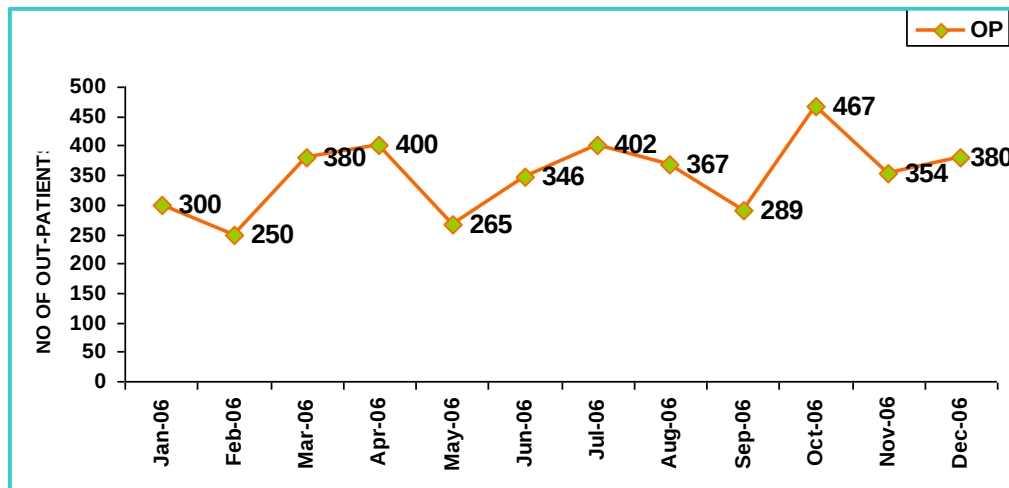
- 1. Histogram**
- 2. Frequency polygon**
- 3. Frequency curve**
- 4. Line graph**
- 5. Scatter or dot diagram**

Presentation of qualitative data is through diagrams, the common diagrams in use are:-

- 1. Bar diagram**
- 2. Pie/sector diagram**
- 3. Pictogram or picture diagram**
- 4. Map diagram or spot map**

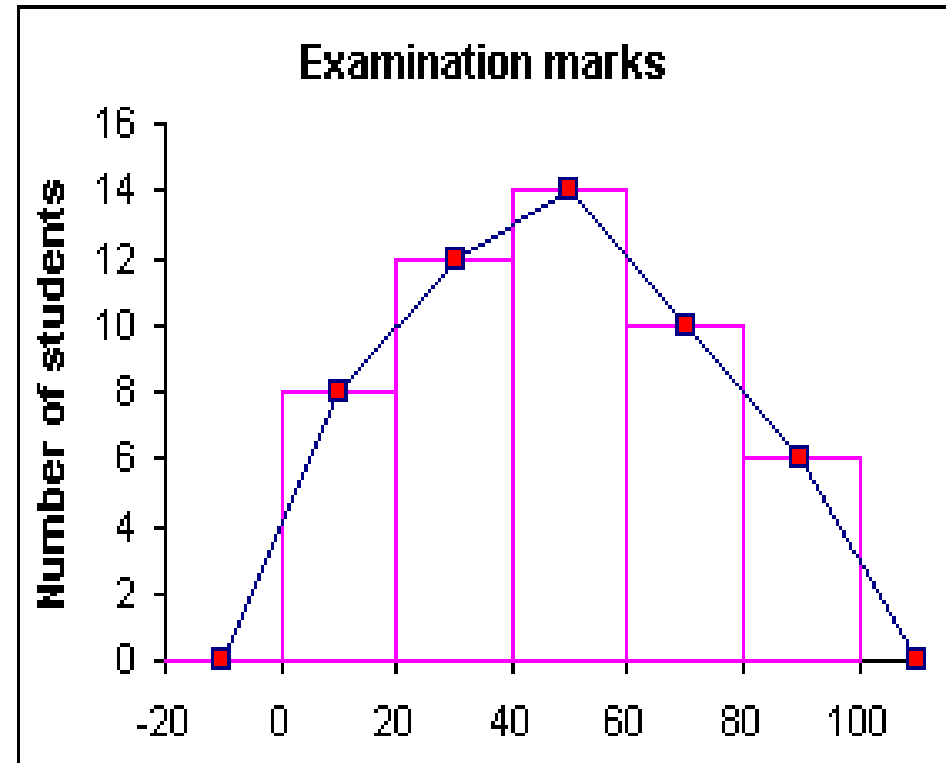
Line diagram:

- This diagram is useful to study changes of values in the variable overtime.
- Simplest type of diagram.
- On the X axis the time such as hours, days, weeks, months or years are represented.
- The value of any quantity pertaining to this is represented along the Y axis.



Frequency polygon:

1. The most commonly used graphic device to illustrate statistical distribution.
2. Used to represent frequency distribution of quantitative data.
3. Useful to compare 2 or more frequency distributions.



- A frequency polygon is a variation of a histogram, in which the bars are replaced by lines connecting the midpoints of the tops of the bars.
- Advocates of the frequency polygon argue that the purpose of a histogram is to show the shape of the data distribution and removing the bars makes the shape clearer and smoother.

Construction of frequency polygon:

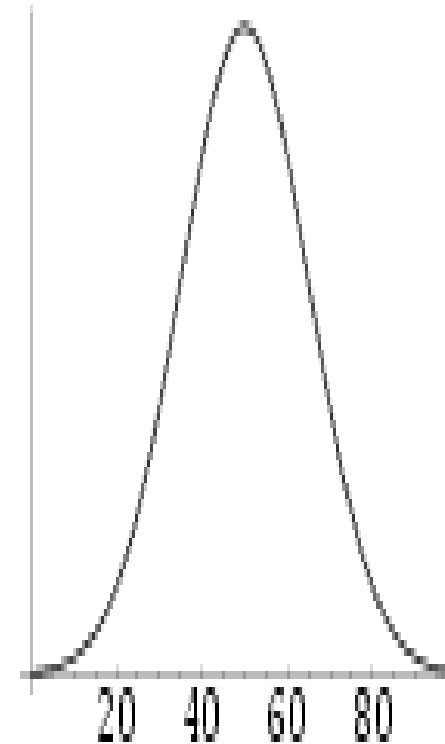
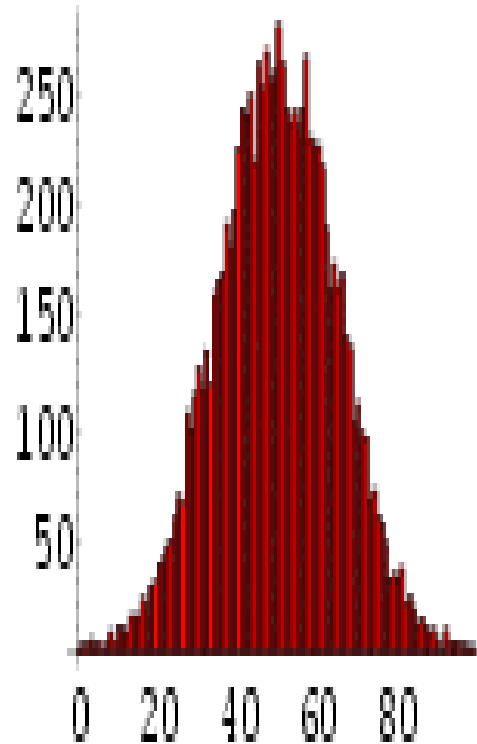
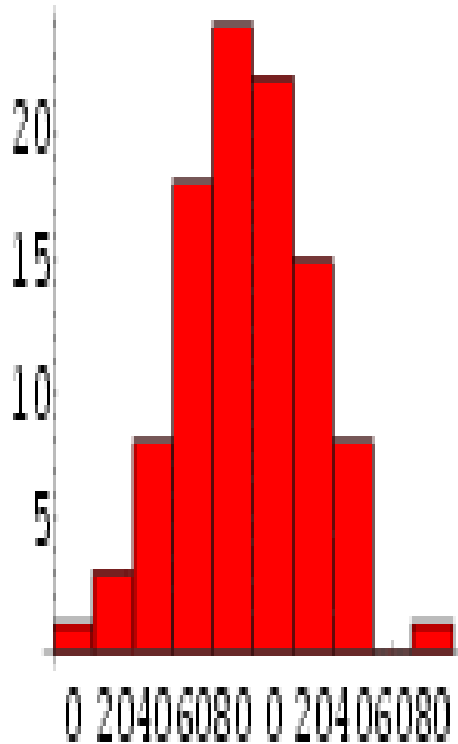
1. To draw a frequency polygon, A point is marked in the midpoint of the class interval, corresponding to the frequency.
2. Then these points are connected by straight lines.
3. The first point and last point are joined to the midpoint of previous and next class respectively.
4. Rather than leaving the graph suspended in space , we assume that there is another interval above and below which is having frequency of 0 .
5. so the midpoint of group are assume having frequency of 0.the group is now allowed to meet x-axis on both the ends.

Advantages of frequency polygon:

- It is very easy to construct and very easy to interpret.
- It is useful in portraying more than two distributions on the same graph paper with different colors. So it is very useful to compare 2 or more than 2 distributions.

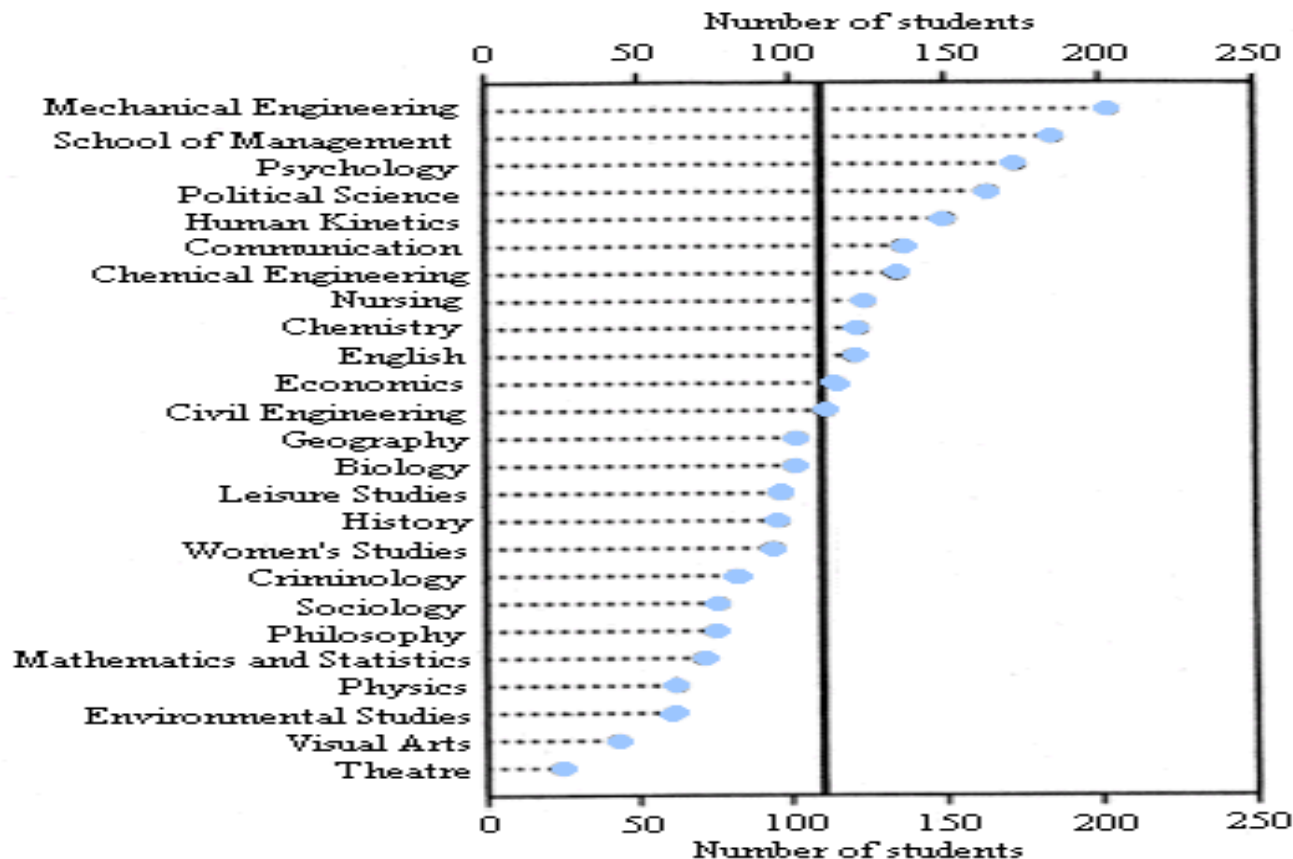
Frequency curve:-

When the number of observations are very large and groups are more (ie; small class intervals) the frequency polygon tends to lose its angulation and it forms a smooth curve known as frequency curve.



Scatter diagram or dot diagram:

1. It is a graphic presentation of data.
2. It is used to show the nature of correlation between 2 variables

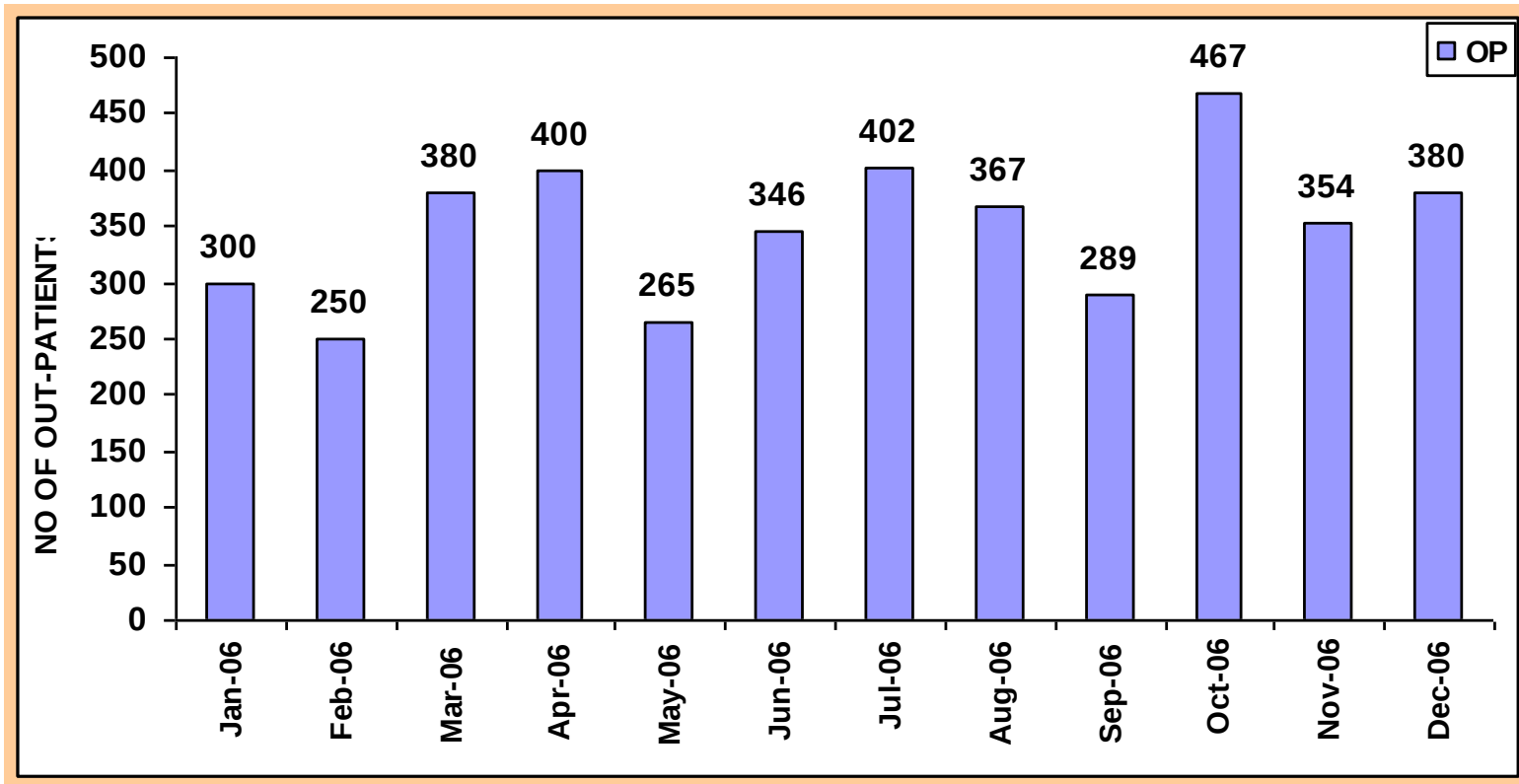


Bar diagram

1. This diagram is used to represent qualitative data.
2. It represent only one variable.
3. The width of the bar remains the same and only the length varies according to the frequency in each category.
4. There are 3 types of bars:
 - a) simple bar
 - b) multiple bar
 - c) component bar diagram

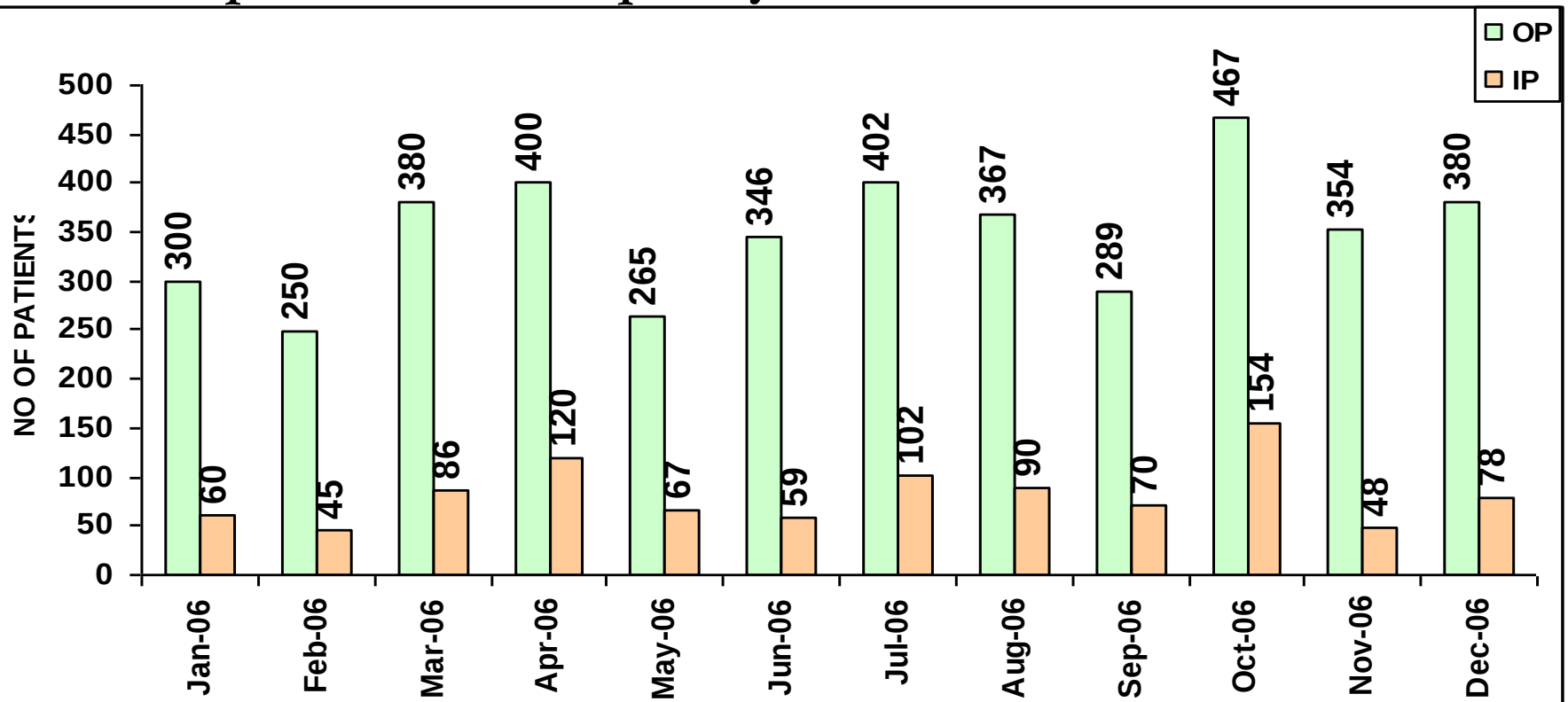
- **Simple bar:**

The limitation of this method is that they can represent only on the classification and hence cannot be used for comparison



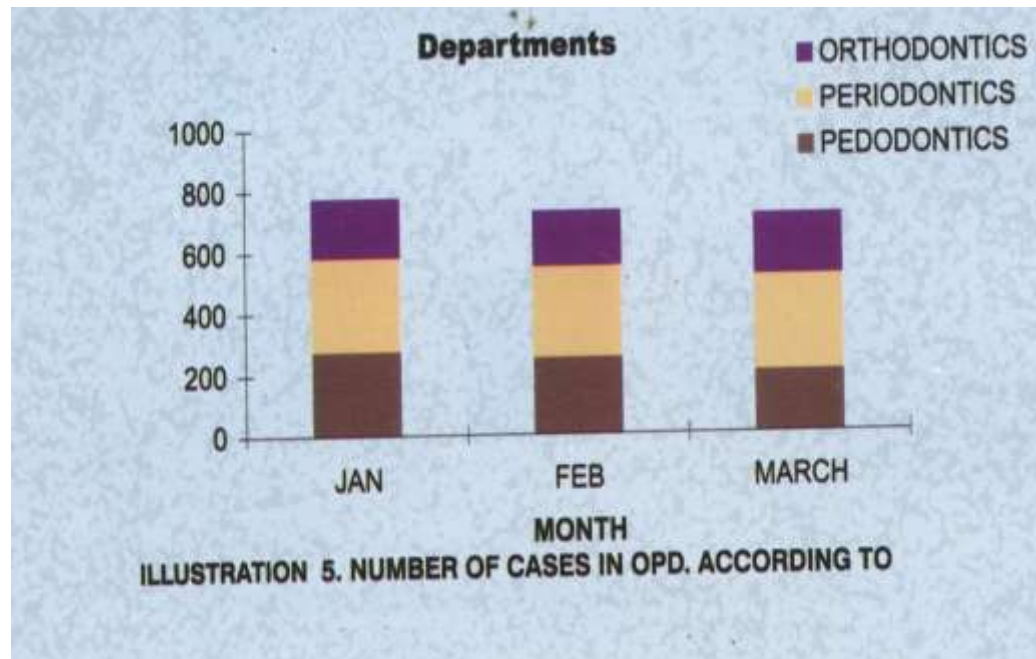
- **Multiple bar:**

1. This diagram is used to compare qualitative data with respect to a single variable.
2. This diagram is similar to the bar diagram except that for each category of the variable we have a set of bars of the same width corresponding to the different section without any gap in between the width and the length corresponds to the frequency.



Component bar diagram:

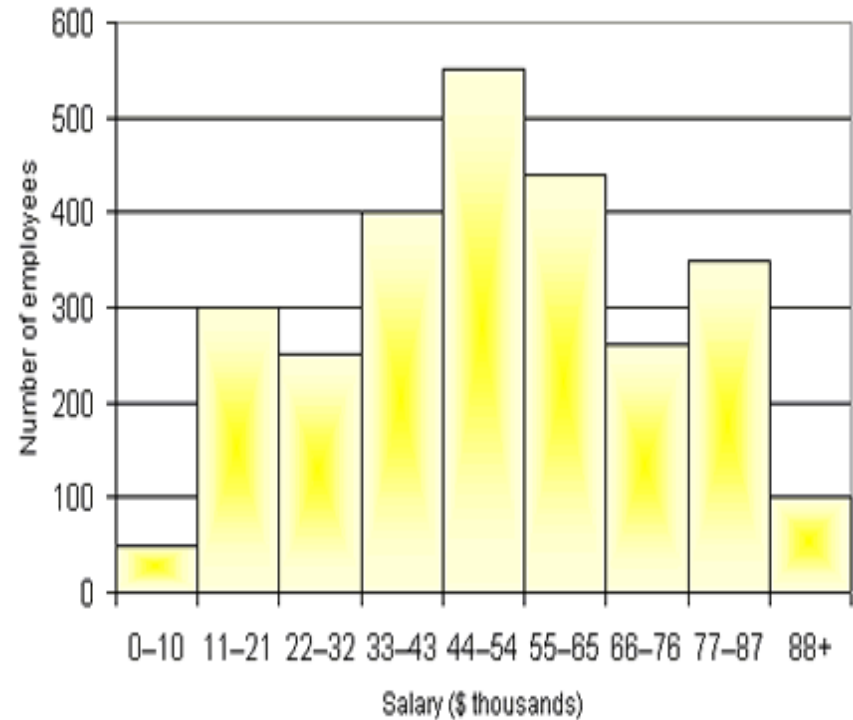
- This diagram is used to represent qualitative data.
- It is desired to represent both the no of cases in major groups as well as the subgroups simultaneously



Histogram:

this diagram is used to depict quantitative data

1. It is a bar diagram without gap between bars.
2. If we draw frequencies of each group or class intervals in the form of columns or rectangles such a diagram is called histogram.
3. It represents a frequency distribution.



The histogram is constructed as follows:

- 1. On the X axis the size of the observation is marked.**
- 2. Starting from 0 the limit of each class interval is marked, the width corresponding to the width of the class interval in the frequency distribution.**
- 3. On the Y axis the frequencies are marked**
- 4. A rectangle is drawn above each class interval with height proportional to the frequency of that interval.**

Advantages of Histogram:

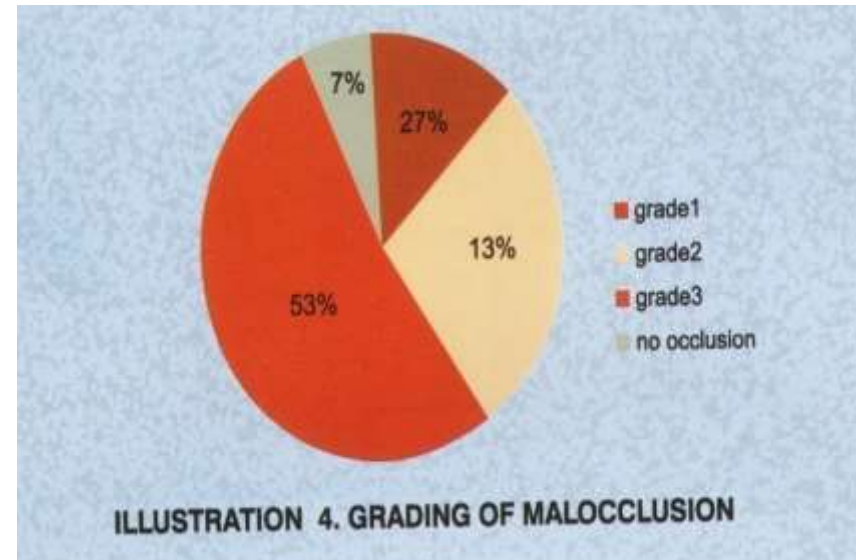
Easy to understand

Disadvantages of Histogram:

- Only 1 histogram can be placed at a time.
- More time consuming to construct than a frequency polygon.

Pie diagram:

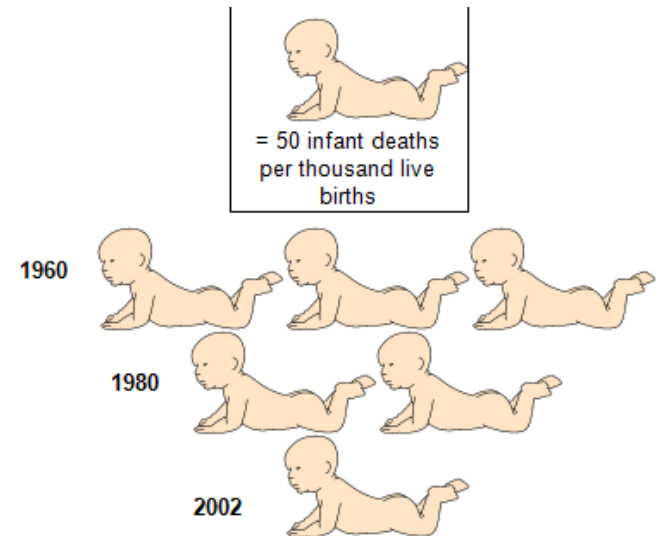
1. These are popularly used to show percentage breakdowns for qualitative data.
2. It is so called because the entire graph looks like a pie and its components represent slices cut from a pie.
3. A circle is divided into different sectors corresponding to the frequencies of the distribution.
4. Some knowledge of circles and degrees is necessary



5. The total angle at the center of the circle is 360 degrees and it represents the total frequency.
6. After the calculation of angle, segments are drawn in the circle and the segments are shaded with different shades or colors and an index is provided for the shaded colors
7. Cannot be used to represent 2 or more data sets

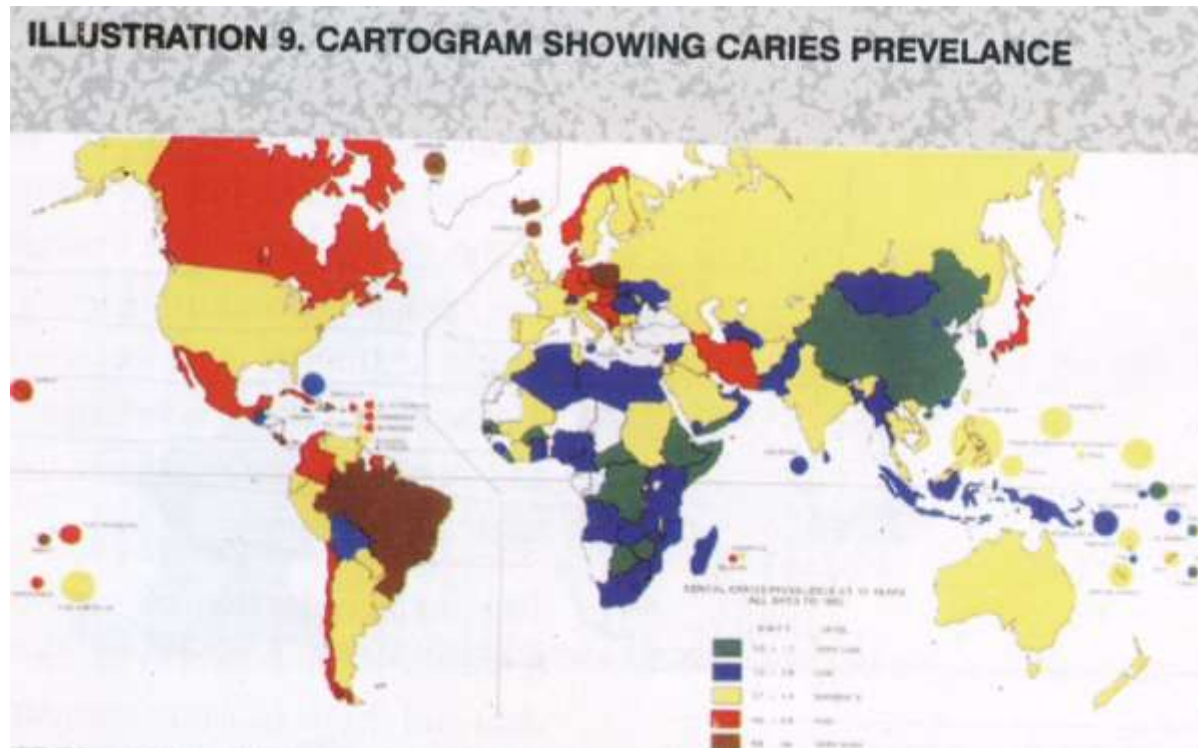
Pictogram

1. Display of data through pictograms was initiated by Dr Otto Neurath in 1923.
2. Data are displayed by the pictures of the items to which the data pertain.
3. A single picture represents a fixed no.
4. They are the least satisfactory type of diagrams.
5. They are inaccurate too.



Map diagram or spot map or cartograms:

1. These maps are used to show geographical distribution of frequencies of a characteristics.



References

1. Kirkwood BR, Sterne JAC. Essentials in Medical Statistics. 2nd edition. Oxford: Blackwell; (2010).
2. Jay S. Kim And Ronald J. Dailey. Biostatistics For Oral Healthcare. Blackwell Publishing Company. 2008
3. C.R Kothari. Research Methodology methods and technologies. 2nd edition. New age international private ltd publishers; 2004. reprint 2007

