Descriptive Statistics

Methods of data, organization and presentation

Session Objectives

By the end of this session the students will be able to:

- List the different methods of data organization
- List scales of measurement
- Use graphs and charts for data presentation

Descriptive Statistics

- Techniques used to organize and summarize a set of data in a concise way.
  - *Organization of data*
  - *Summarization of data*
  - *Presentation of data*
- Numbers that have not been summarized and organized are called **raw data**.

Data presentations

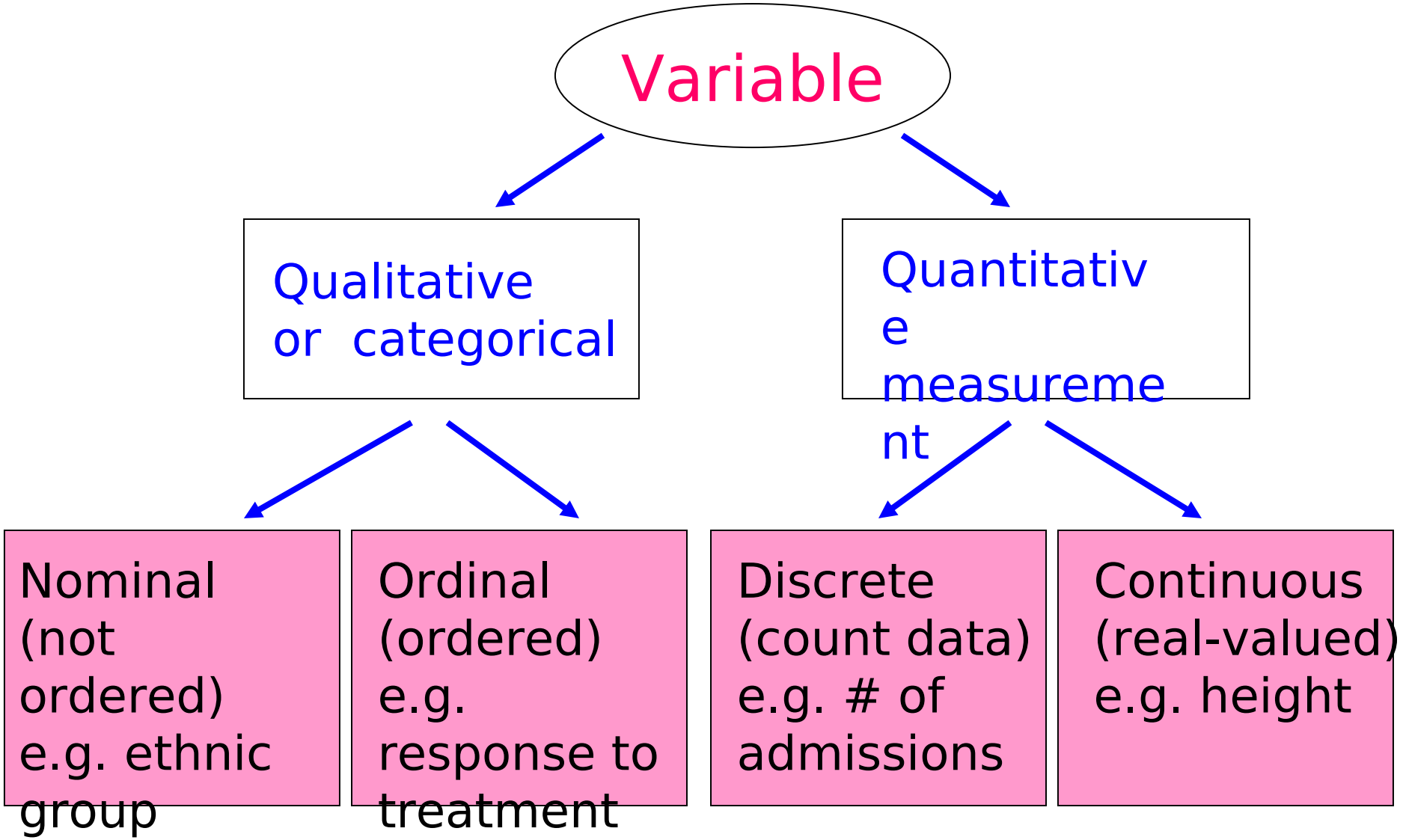
- Tables
- Graphs
- Numerical summary measures
  - Measures of central tendency
  - Measures of variability

- Before summarization and organization, we need to know the types of variables and measurement scales of our data.

- Before displaying or analyzing data, classify the variables into their different types.

Variable

- **Variable**: A characteristic which takes different values in different persons, places, or things.
- Any aspect of an individual or object that is measured (e.g., BP) or recorded (e.g., age, sex) and takes any value.
- There may be one variable in a study or many.
- E.g., A study of treatment outcome of TB


Variable
Qualitative or categorical
Quantitative measurement
Nominal (not ordered) e.g. ethnic group
Ordinal (ordered) e.g. response to treatment
Discrete (count data) e.g. # of admissions
Continuous (real-valued) e.g. height

- **Variables can be broadly classified into:**
  - ***Categorical (or Qualitative) or***
  - ***Quantitative (or numerical variables).***

- **Categorical variable**: A variable or characteristic which can not be measured in quantitative form but can only be sorted by name or categories

- Not able to be measured as we measure height or weight

- The notion of magnitude is absent or implicit.

- **Quantitative variable**: A variable that can be measured (or counted) and expressed numerically.

- Height, wt, # of children, etc.

- Has the notion of magnitude.

Quantitative variable is divided into two:

1. Discrete: It can only have a limited number of discrete values (usually whole numbers).

- *E.g., the number of episodes of diarrhoea a child has had in a year. You can't have 12.5 episodes of diarrhoea*

- Characterized by gaps or interruptions in the values (integers).
- Both the order and magnitude of the values matter.
- The values aren't just labels, but are actual measurable quantities.

2. Continuous variable: It can have an infinite number of possible values in any given interval.

- Both the magnitude and the order of the values matter
- Does not possess the gaps or interruptions
- Weight is continuous since it can take on any number of values (e.g., 34.575 Kg).

Scales of measurement

- All measurements are not the same.
- Measuring weight = eg. 40kg
- Measuring the status of a patient on scale = “improved”, “stable”, “not improved”.
- There are **four** types of scales of measurement.

1. **Nominal scale**:

- The simplest type of data, in which the values fall into unordered categories or classes
- Consists of “naming” observations or classifying them into various mutually exclusive and collectively exhaustive categories
- Uses names, labels, or symbols to assign each measurement.
  - *Examples: Blood type, sex, race, marital status, etc.*

Example of nominal Scale:

Race/Ethnicity:

1. Black
2. White
3. Latino
4. Other

- The numbers have NO meaning
- They are labels only

- If nominal data can take on only two possible values, they are called *dichotomous or binary*.
- So sex is not just nominal, it is dichotomous (male or female).
- Yes/no questions
  - E.g., cured from TB at 6 months of Rx

2. Ordinal scale:

- Assigns each measurement to one of a limited number of categories that are ranked in terms of order.
- Although non-numerical, can be considered to have a natural ordering
  - *Examples: Patient status, cancer stages, social class, etc.*

Example of ordinal scale:

- **Pain level**:
  1. None
  2. Mild
  3. Moderate
  4. Severe

- The numbers have LIMITED meaning 4>3>2>1 is all we know apart from their utility as labels

3. **Interval scale**:

- Measured on a continuum and differences between any two numbers on a scale are of known size.

  Example: Temp. in °F on 4 consecutive days

Days:	A	B	C	D
Temp. °F:	50	55	60	100

  For these data, not only is day A with 50° cooler than day D with 65°, but is 15° cooler.

- It has no true zero point. “0” is arbitrarily chosen and doesn’t reflect the absence of temp.

4. Ratio scale:

- Measurement begins at a true zero point and the scale has equal space.
  - Examples: Height, age, weight, BP, etc.
- *Note on meaningfulness of "ratio"-*
  - Someone who weighs 80 kg is two times as heavy as someone else who weighs 40 kg. This is true even if weight had been measured in other measurements.

Exercises:

Give the correct scales of measurement for each variable

1. Blood group
2. Temperature (Celsius)
3. Hair colour
4. Job satisfaction index (1-5)
5. Serum uric acid (mg/100ml)
6. The average weight gain of 1-year old dogs with a special diet supplement was 950 grams last month
7. Ethnic group

Methods of Data Organization and Presentation

- The actual summarization and organization of data starts from frequency distribution.

- **Frequency distribution**: ***A table which has a list of each of the possible values that the data can assume along with the number of times each value occurs.***

- For **nominal and ordinal data**, frequency distributions are often used as a summary.
- Example

Gender	Frequency	Relative Frequency
Male	30	65.2%
Female	16	34.8%
Total	46	100.0%

- The % of times that each value occurs, or the **relative frequency**, is often listed
- Tables make it easier to see how the data are distributed

- For both **discrete and continuous** data, the values are grouped into non-overlapping intervals, usually of equal width.

Class Interval	Frequency
30–39	11
40–49	46
50–59	70
60–69	45
70–79	16
80–89	1
Total	189

a) **Qualitative variable**: Count the number of cases in each category.

- **Example1**: The intensive care unit type of 25 patients entering ICU at a given hospital:
  1. **Medical**
  2. **Surgical**
  3. **Cardiac**
  4. **Other**

ICU Type	Frequency (How often)	Relative Frequency (Proportionately often)
Medical	12	0.48
Surgical	6	0.24
Cardiac	5	0.20
Other	2	0.08
Total	25	1.00

Example 2:

A study was conducted to assess the characteristics of a group of 234 smokers by collecting data on gender and other variables.

Gender, 1 = male, 2 = female

Gender	Frequency (n)	Relative Frequency
Male (1)	110	47.0%
Female (2)	124	53.0%
Total	234	100%

b) **Quantitative variable**:

- Select a set of continuous, non-overlapping intervals such that each value can be placed in one, and only one, of the intervals.
- The first consideration is how many intervals to include

Age	Frequency
15	1
19	2
29	1
31	1
34	1
39	1
52	2
53	2
45	2
65	1
71	1
74	1
75	2
76	2
77	1
78	1
79	1
84	1
85	1

For a continuous variable (e.g. – age), the frequency distribution of the individual ages is not so interesting.

Age Interval	Frequency
10-19	3
20-29	1
30-39	3
40-49	0
50-59	6
60-69	1
70-79	9
80-89	2
TOTAL	25

- **We “see more” in frequencies of age values in “groupings”. Here, 10 year groupings make sense.**
- **Grouped data frequency distribution**

To determine the number of class intervals and the corresponding width, we may use:

Sturge's rule: $K = 1 + 3.322(\log n)$

$$W = \frac{L - S}{K}$$

where

K = number of class intervals n = no. of observations

W = width of the class interval L = the largest value

S = the smallest value

Example:

– free time (hours) per week for 40 college students:

23 24 18 14 20 36 24 26 23 21 16 15 19 20 22 14 13 10 19 27 29 22 38 28 34 32 23 19 21 31 16 28 19 18 12 27 15 21 25 16

K = 1 + 3.322 (log40) = 6.32 ≈ 6

Maximum value = 38, Minimum value = 10

Width = (38-10)/6 = 4.66 ≈ 5

Time (Hours)	Frequency	Relative Frequency	Cumulative Relative Frequency
10-14	5	0.125	0.125
15-19	11	0.275	0.400
20-24	12	0.300	0.700
25-29	7	0.175	0.875
30-34	3	0.075	0.950
35-39	2	0.050	1.00
Total	40	1.00	

- **Cumulative frequencies: When frequencies of two or more classes are added.**

- **Cumulative relative frequency: The percentage of the total number of observations that have a value either in that interval or below it.**

- **Mid-point: The value of the interval which lies midway between the lower and the upper limits of a class.**

- **True limits**: Are those limits that make an interval of a continuous variable continuous in both directions

- Used for smoothening of the class intervals

- Subtract 0.5 from the lower and add it to the upper limit

Time (Hours)	True limit	Mid-point	Frequency
10-14	9.5 – 14.5	12	5
15-19	14.5 – 19.5	17	11
20-24		22	12
25-29	19.5 – 24.5	27	7
30-34		32	3
35-39	24.5 – 29.5	37	2
	29.5 – 34.5		

Simple Frequency Distribution

- Primary and secondary cases of syphilis morbidity by age, 1989

Age group (years)	Cases	
	Number	Percent
0-14	230	0.5
15-19	4378	10.0
20-24	10405	23.6
25-29	9610	21.8
30-34	8648	19.6
35-44	6901	15.7
45-54	2631	6.0
>44	1278	2.9
Total	44081	100

Two Variable Table

- Primary and secondary cases of syphilis morbidity by age and sex, 1989

Age group (years)	Number of cases		
	Male	Female	Total
0-14	40	190	230
15-19	1710	2668	4378
20-24	5120	5285	10405
25-29	5301	4306	9610
30-34	5537	3111	8648
35-44	5004	1897	6901
45-54	2144	487	2631
>44	1147	131	1278
Total	26006	18075	44081

Tables can also be used to present more than three or more variables.

Variable	Frequency (n)	Percent
Sex		
Male		
Female		
Age (yrs)		
15-19		
20-24		
25-29		
Religion		
Christian		
Muslim		
Occupation		
Student		
Farmer		
Merchant		

Guidelines for constructing tables

- **Keep them simple,**
- **Limit the number of variables to three or less,**
- **All tables should be self-explanatory,**
- **Include clear title telling what, when and where,**
- **Clearly label the rows and columns,**
- **State clearly the unit of measurement used,**
- **Explain codes and abbreviations in the foot-note,**
- **Show totals,**
- **If data is not original, indicate the source in foot-note.**

Diagrammatic Representation

- Pictorial representations of numerical data

Importance of diagrammatic representation:

1. Diagrams have greater attraction than mere figures.
2. They give quick overall impression of the data.
3. They have great memorizing value than mere figures.
4. They facilitate comparison
5. Used to understand patterns and trends

Specific types of graphs include:

- Bar graph
- Pie chart

→ **Nominal, ordinal data**

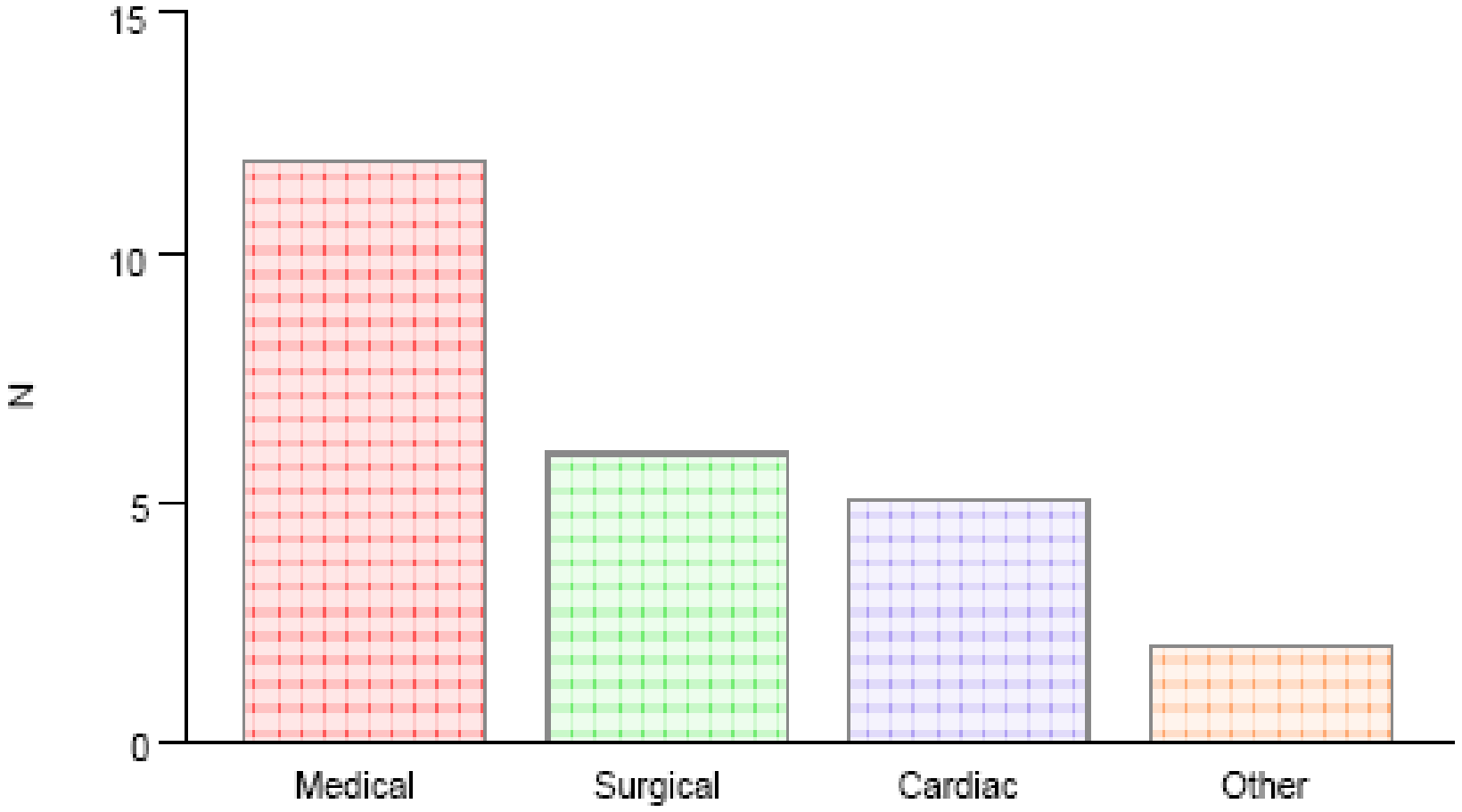
- Histogram
- Stem-and-leaf plot
- Box plot
- Scatter plot
- Line graph
- Others

→ **Quantitative data**

1. Bar charts (or graphs)

- Categories are listed on the horizontal axis (X-axis)
- Frequencies or relative frequencies are represented on the Y-axis (ordinate)
- The height of each bar is proportional to the frequency or relative frequency of observations in that category

Bar chart for the type of ICU for 25 patients



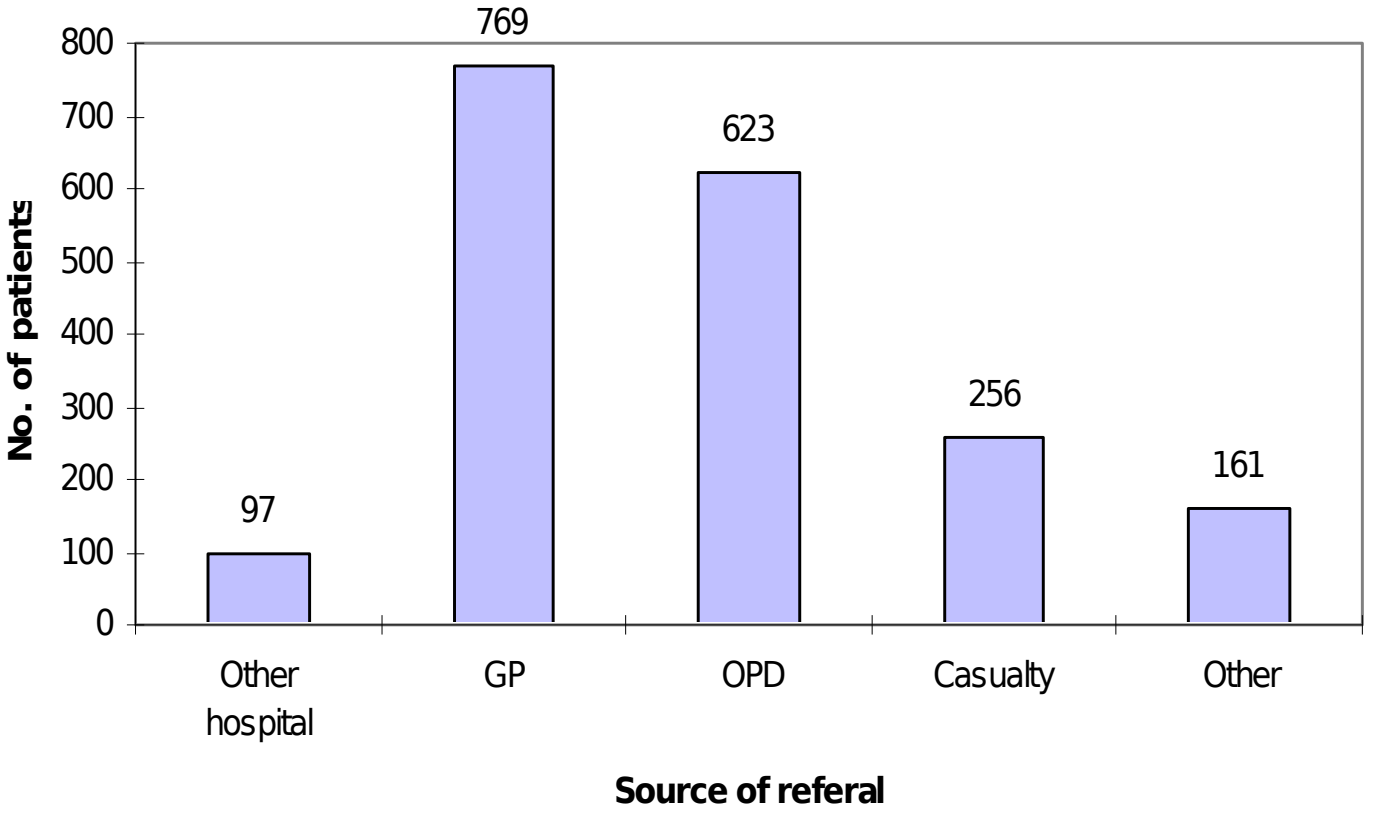
Method of constructing bar chart

- All the bars must have equal width
- The bars are not joined together (leave space between bars)
- The different bars should be separated by equal distances
- All the bars should rest on the same line called the base
- Label both axes clearly

Example: Construct a bar chart for the following data.

Distribution of patients in hospital by source of referral

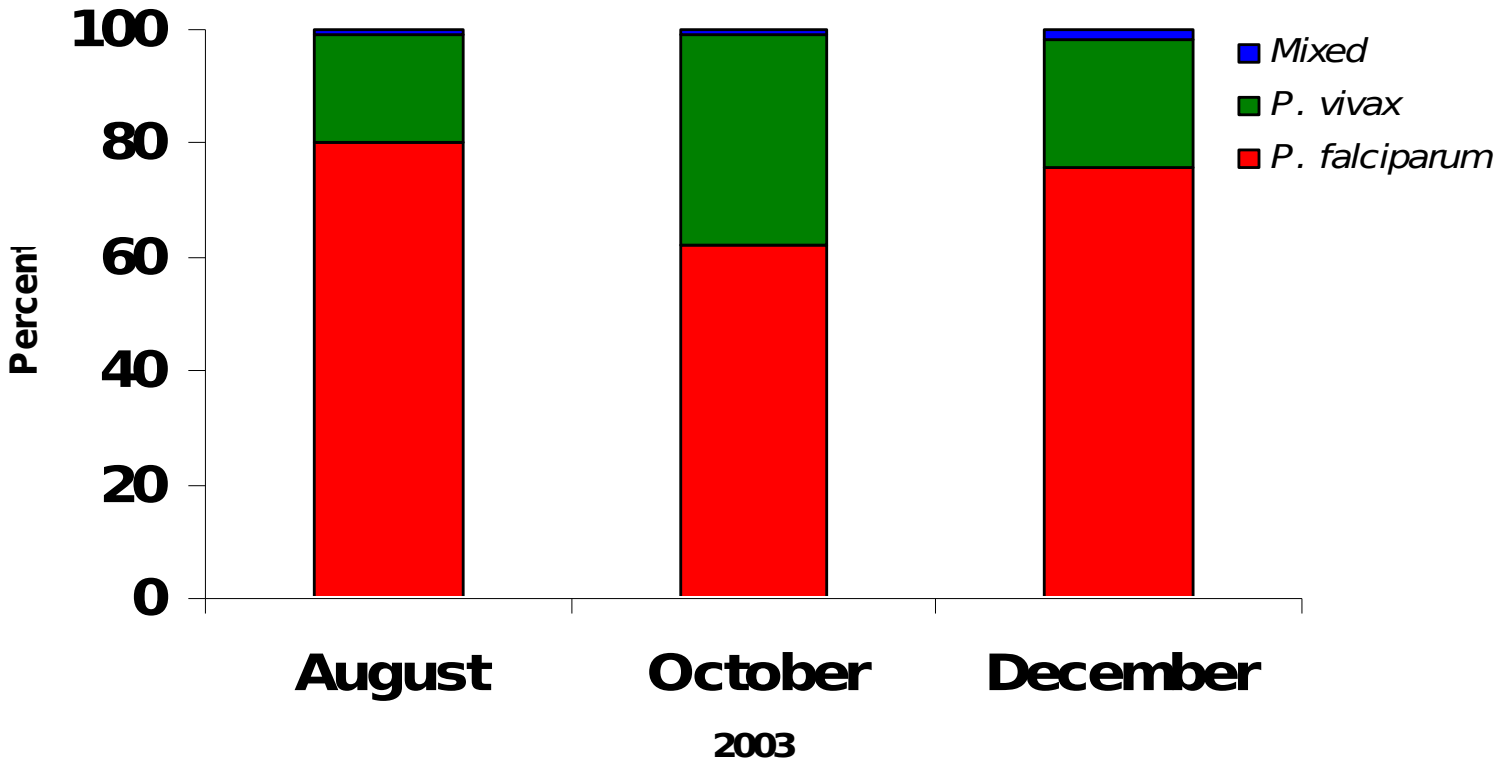
Source of referral	No. of patients	Relative freq.
Other hospital	97	5.1
General practitioner	769	40.3
Out-patient department	623	32.7
Casualty	256	13.4
Other	161	8.5
Total	1 906	100.0


Distribution of patients in hopital X by source of referal, 1999
No. of patients
800
700
600
500
400
300
200
100
0
769
623
256
161
97
Other hospital
GP
OPD
Casualty
Other
Source of referal

2. Sub-divided bar chart

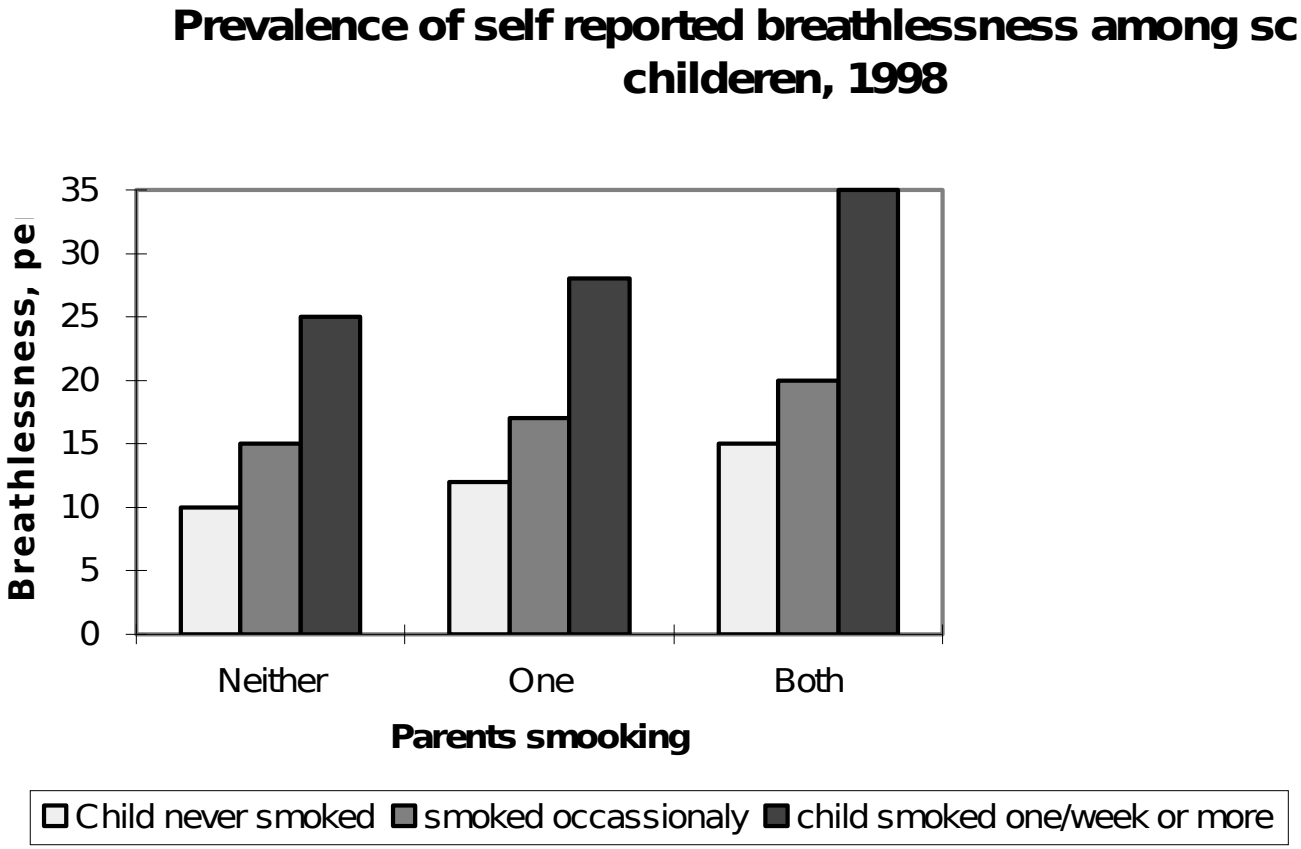
- If there are different quantities forming the sub-divisions of the totals, simple bars may be sub-divided in the ratio of the various sub-divisions to exhibit the relationship of the parts to the whole.
- The order in which the components are shown in a “bar” is followed in all bars used in the diagram.

Example: Plasmodium species distribution for confirmed malaria cases, Zeway, 2003



3. Multiple bar graph

- Bar charts can be used to represent the relationships among more than two variables.
- The following figure shows the relationship between children’s reports of breathlessness and cigarette smoking by themselves and their parents.



We can see from the graph quickly that the prevalence of the symptoms increases both with the child's smoking and with that of their parents.

There's no reason why the bar chart can't be plotted horizontally instead of vertically.

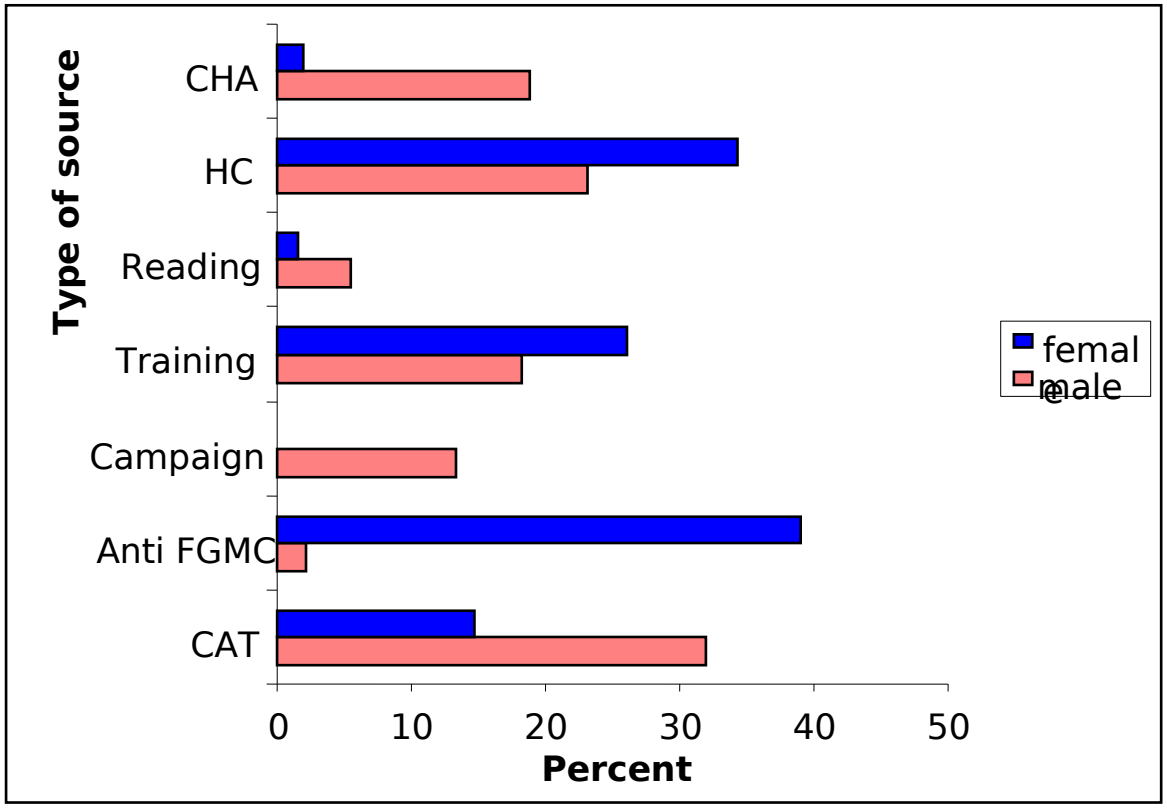


Figure 1. Source of information on the complications of FGM and participation in RH programs, Jijiga, 2004*. * *FGMC = female genital mutilation committee; CAT= community action team; HC = health centre; CHA= community health agent*

4. Pie chart

- Shows the relative frequency for each category by dividing a circle into sectors, the angles of which are proportional to the relative frequency.
- Used for a single categorical variable
- Use percentage distributions

Relative Frequency	Size of Wedge, in degrees
0.50	50% of 360 = 180 degrees
0.25	25% of 360 = 90 degrees
p	(p) x (100%) x 360 degrees

Steps to construct a pie-chart

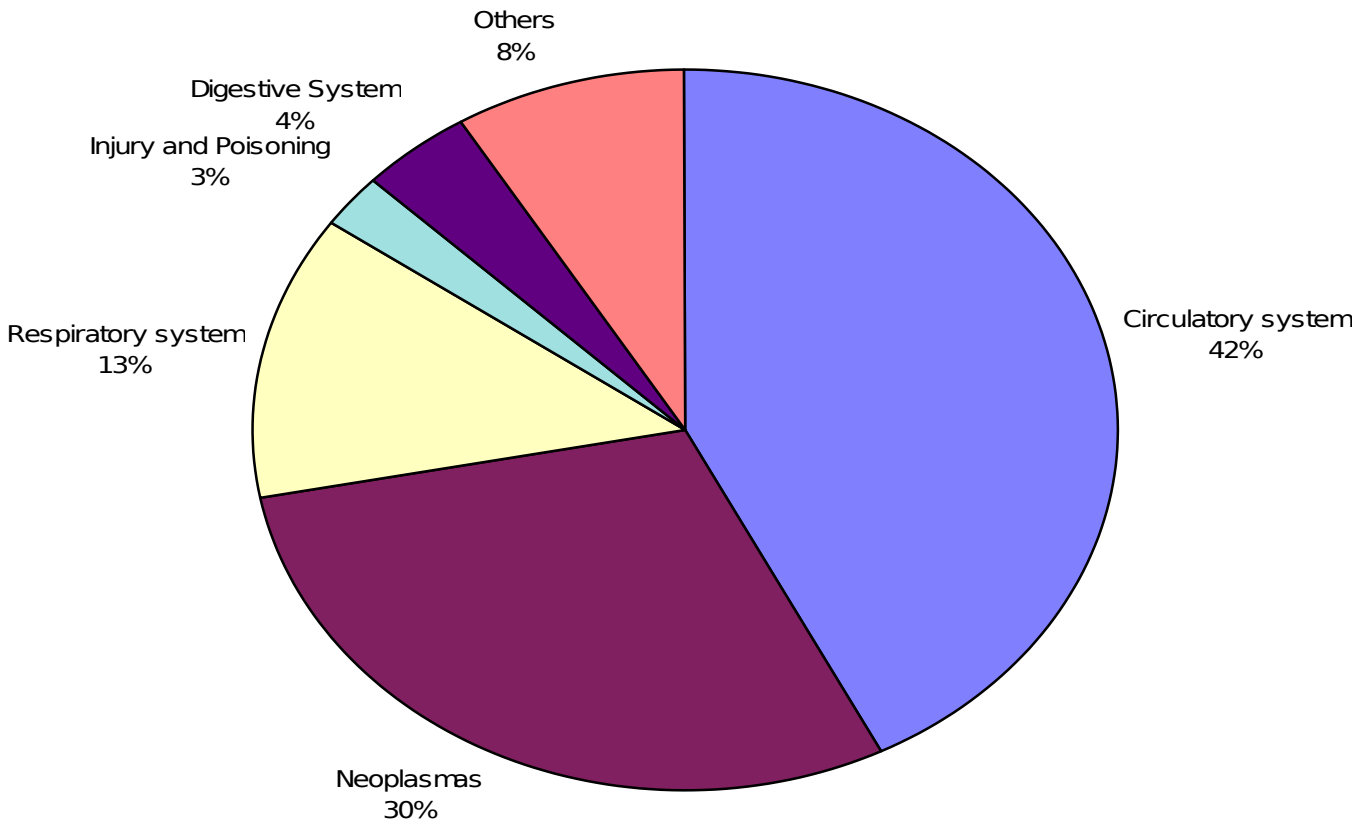
- Construct a frequency table
- Change the frequency into percentage (P)
- Change the percentages into degrees, where: degree = Percentage X 360^{o}
- Draw a circle and divide it accordingly

Example: Distribution of deaths for females, in England

and Wales, 1989.

Cause of death	No. of death
Circulatory system	100 000
Neoplasm	70 000
Respiratory system	30 000
Injury and poisoning	6 000
Digestive system	10 000
Others	20 000
Total	236 000

Distribution fo cause of death for females, in England and Wales, 1989



5. Histogram

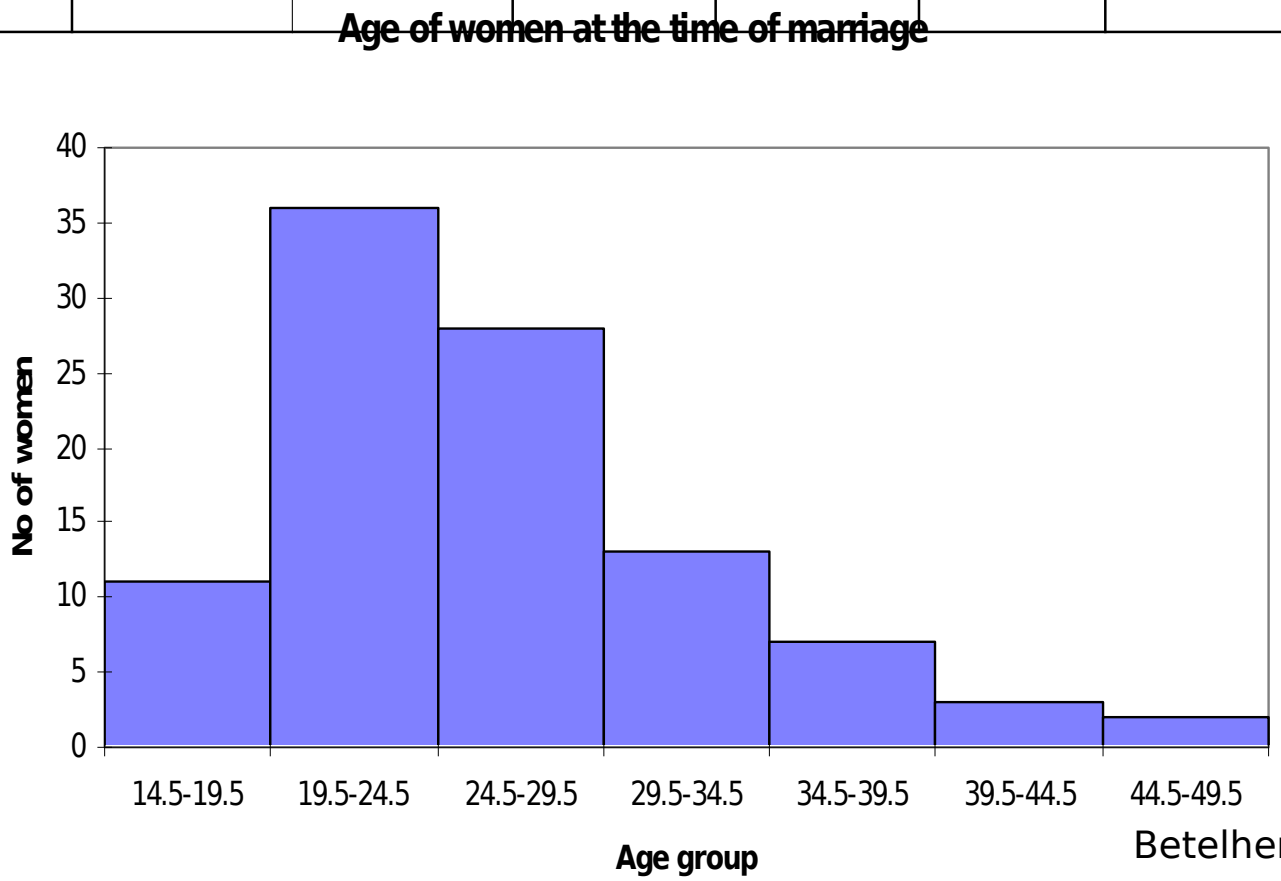
- Histograms are frequency distributions with continuous class intervals that have been turned into graphs.
- To construct a histogram, we draw the interval boundaries on a horizontal line and the frequencies on a vertical line.
- Non-overlapping intervals that cover all of the data values must be used.

- Bars are drawn over the intervals in such a way that the areas of the bars are all proportional in the same way to their interval frequencies.

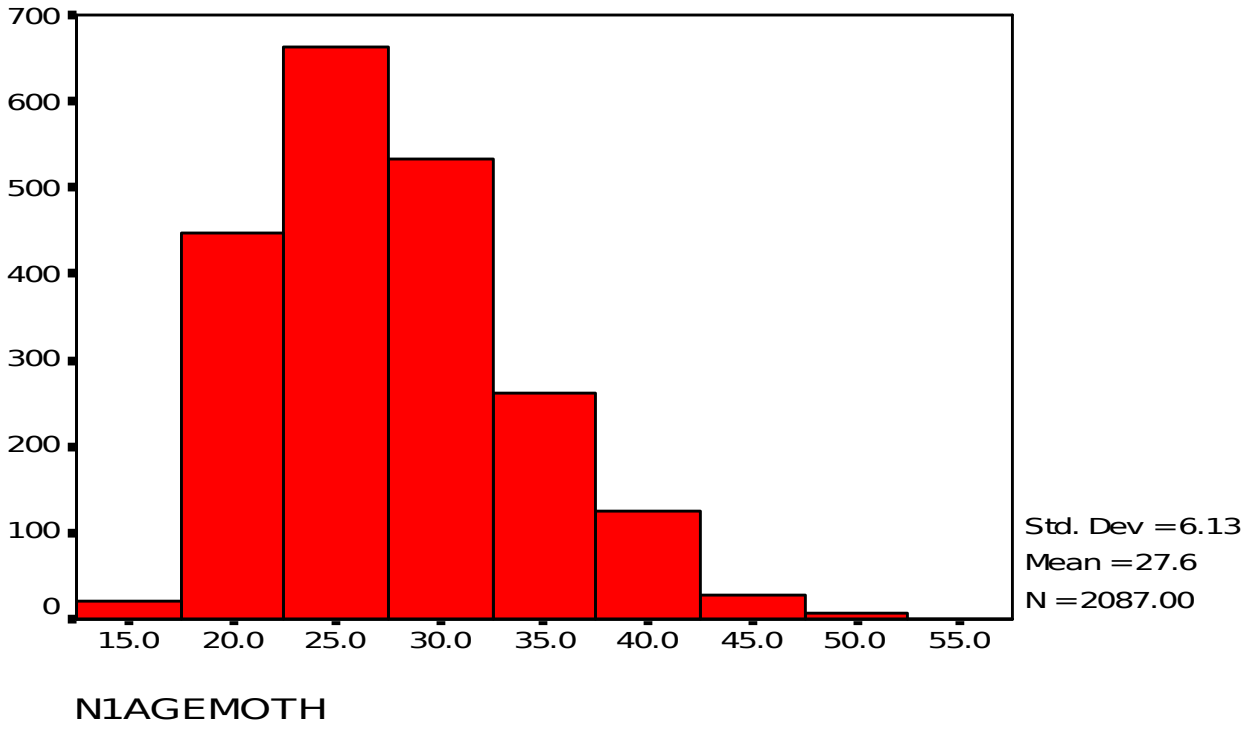
- The area of each bar is proportional to the frequency of observations in the interval

Example: Distribution of the age of women at the time of marriage

Age group	15-19	20-24	25-29	30-34	35-39	40-44	45-49
Number	11	36	28	13	7	3	2



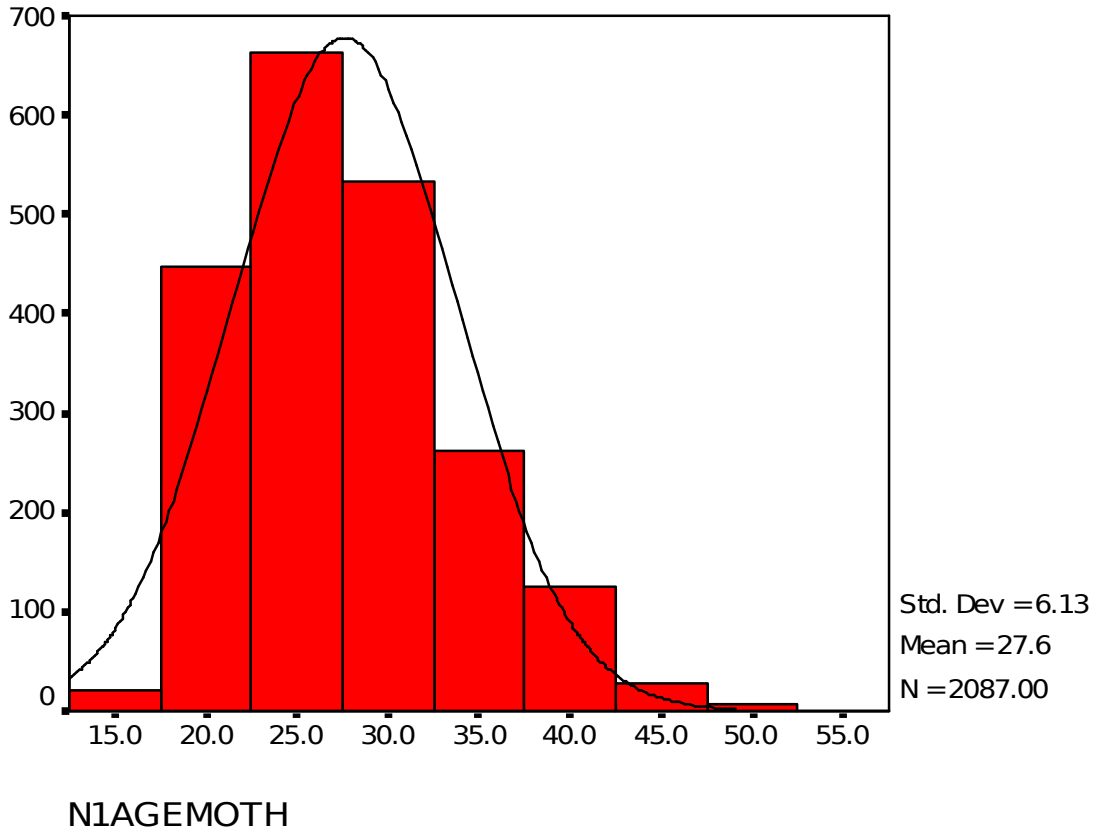
Histogram for the ages of 2087 mothers with <5 children, Adami Tulu, 2003



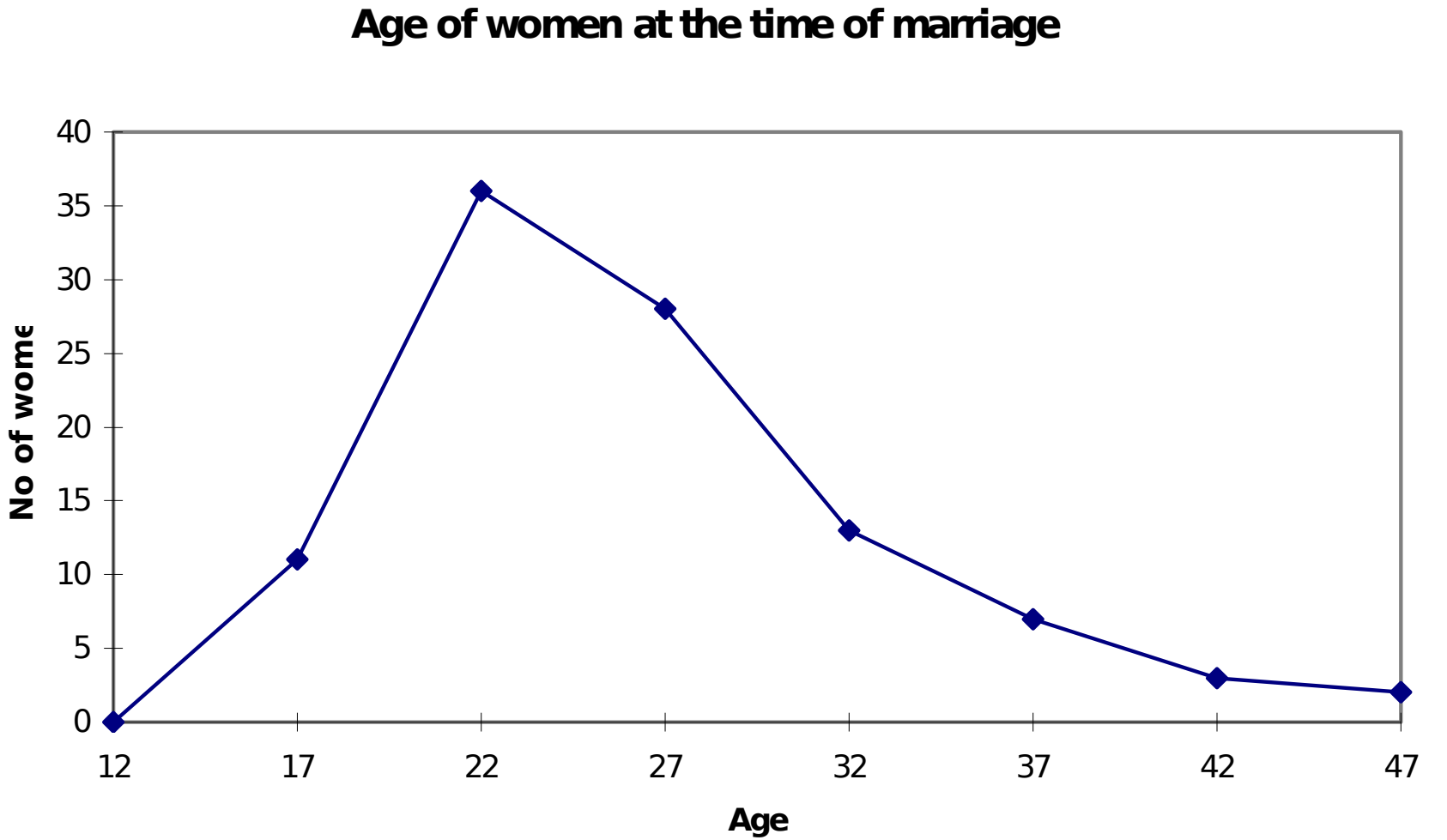
6. **Frequency polygon**

- A frequency distribution can be portrayed graphically in yet another way by means of a frequency polygon.
- To draw a frequency polygon we connect the mid-point of the tops of the cells of the histogram by a straight line.
- The total area under the frequency polygon is equal to the area under the histogram
- Useful when comparing two or more frequency distributions by drawing them on the same diagram

Frequency polygon for the ages of 2087 mothers with <5 children, Adami Tulu, 2003



It can be also drawn without erecting rectangles by joining the top midpoints of the intervals representing the frequency of the classes as follows:



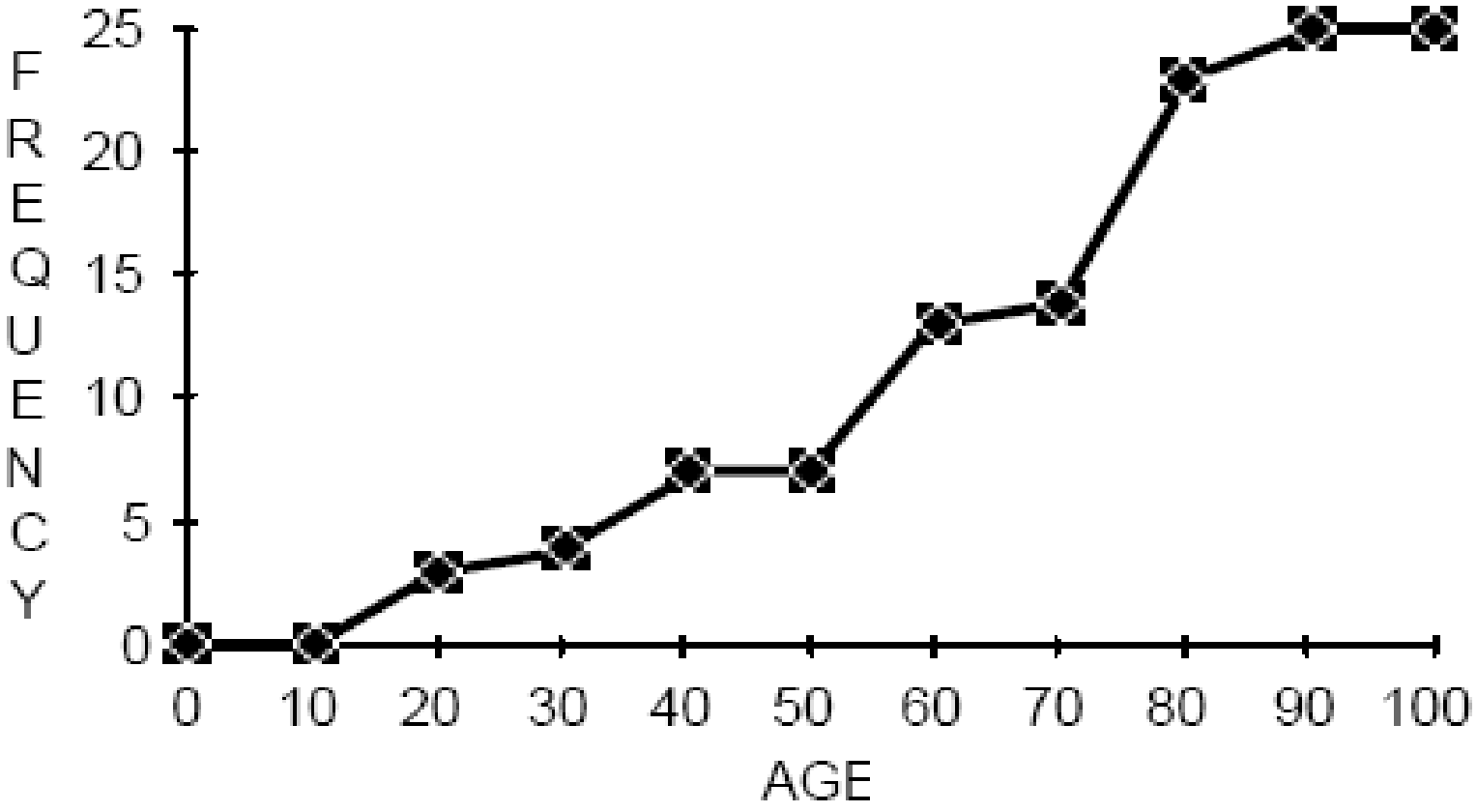
7. Ogive Curve (The Cummulative Frequency Polygon)

- Some times it may be necessary to know the number of items whose values are more or less than a certain amount.
- We may, for example, be interested to know the no. of patients whose weight is <50 Kg or >60 Kg.
- To get this information it is necessary to change the form of the frequency distribution from a ‘simple’ to a ‘cumulative’ distribution.
- Ogive curve turns a cumulative frequency distribution in to graphs.
- Are much more common than frequency polygons

Cumulative Frequency and Cum. Rel. Freq. of Age of 25 ICU Patients

Age Interval	Frequency	Relative Frequency (%)	Cumulative frequency	Cumulative Rel. Freq. (%)
10-19	3	12	3	12
20-29	1	4	4	16
30-39	3	12	7	28
40-49	0	0	7	28
50-59	6	24	13	52
60-69	1	4	14	56
70-79	9	36	23	92
80-89	2	8	25	100
Total	25	100		

Cumulative frequency of 25 ICU patients



8. Scatter plot

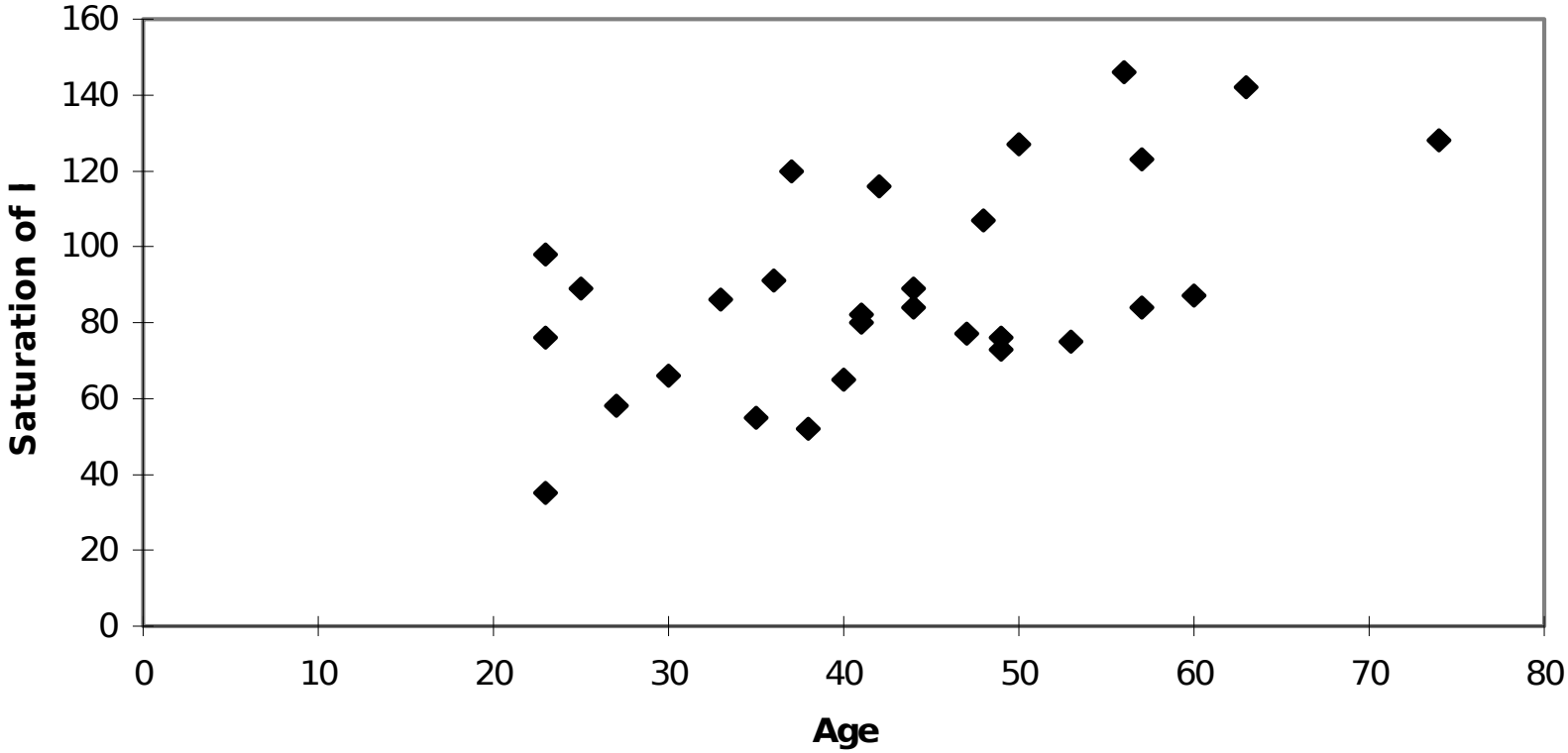
- Most studies in medicine involve measuring more than one characteristic, and graphs displaying the relationship between two characteristics are common in literature.
- When both the variables are qualitative then we can use a multiple bar graph.
- When one of the characteristics is qualitative and the other is quantitative, the data can be displayed in box and whisker plots.

- For two quantitative variables we use bivariate plots (also called scatter plots or scatter diagrams).

- In the study on percentage saturation of bile, information was collected on the age of each patient to see whether a relationship existed between the two measures.

- **A scatter diagram is constructed by drawing X-and Y-axes.**
- **Each point represented by a point or dot(•) represents a pair of values measured for a single study subject.**

Age and percentage saturation of bile for women patients in hospital Z, 1998

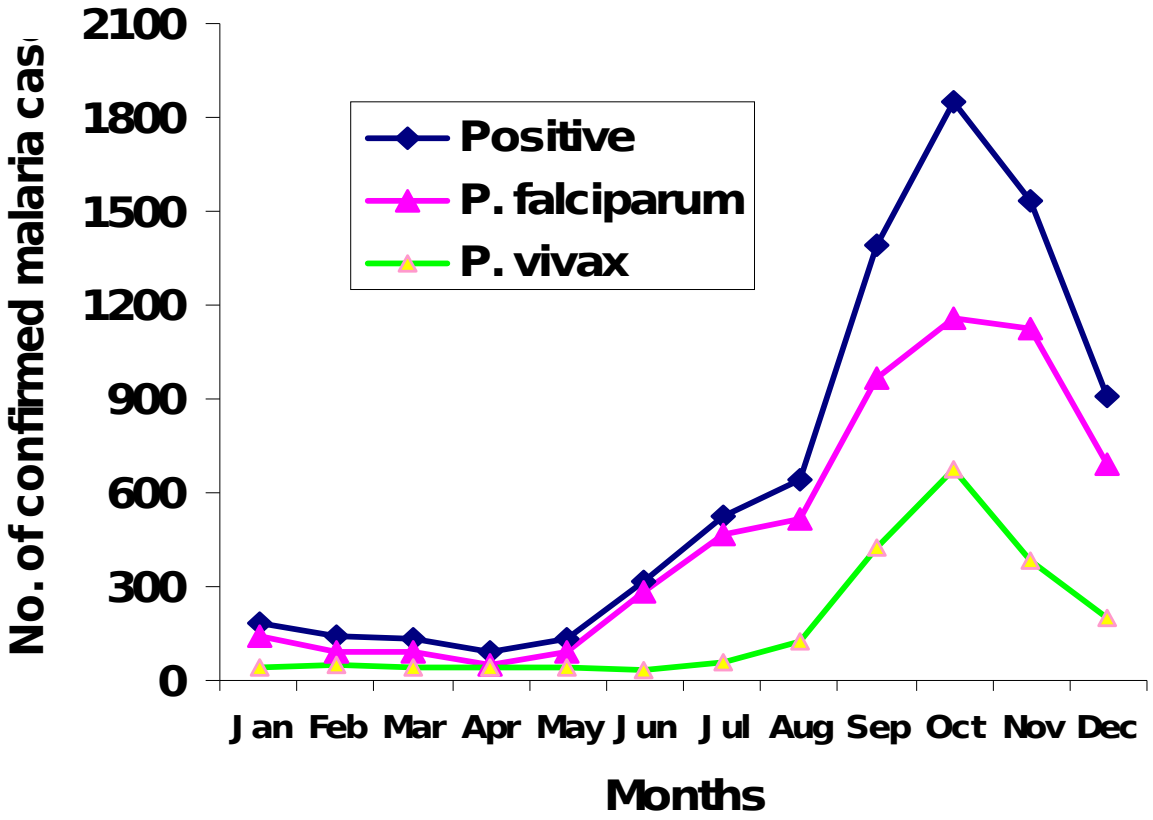


- The graph suggests the possibility of a positive relationship between age and percentage saturation of bile in women.

9. Line graph

- **Useful for assessing the trend of particular situation overtime.**
- **Helps for monitoring the trend of epidemics.**
- **The time, in weeks, months or years, is marked along the horizontal axis, and**
- **Values of the quantity being studied is marked on the vertical axis.**
- **Values for each category are connected by continuous line.**
- **Sometimes two or more graphs are drawn on the same graph taking the same scale so that the plotted graphs are comparable.**

No. of microscopically confirmed malaria cases by species and month at Zeway malaria control unit, 2003



Thank You!