Introduction to Biostatistics

- Prof. Simantini Chattopadhyay
- Assistant Professor
- Department of Economics
- Taki Govt College

Part- I, Zoology Honours

Paper-II, Module-2

Preliminary knowledge of quantification in Biology

Biostatistics

- The application of statistics to a wide range of topics in biology.

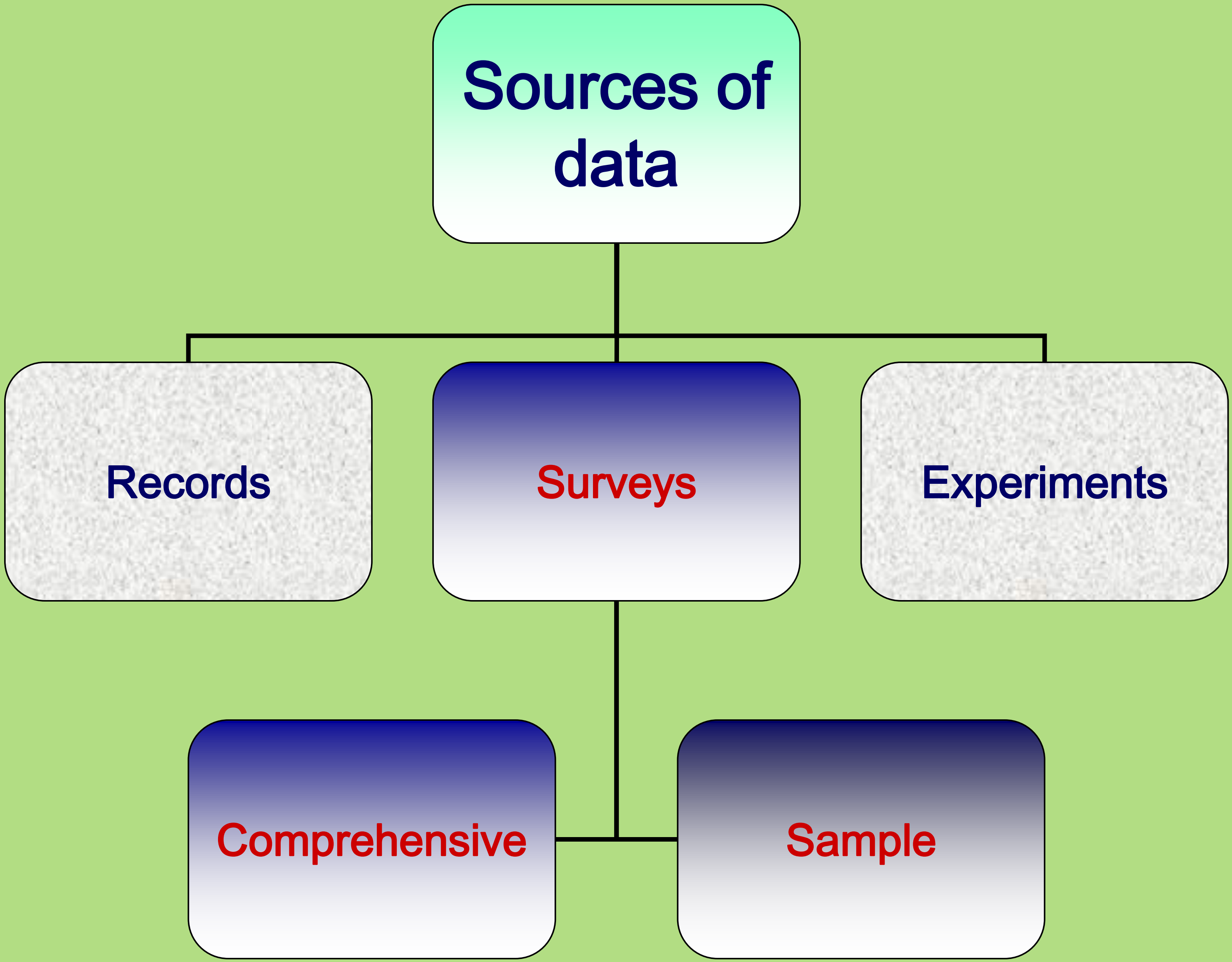
  It is the science which deals with development and application of the most appropriate methods for the:
  - Collection of data.
  - Presentation of the collected data.
  - Analysis and interpretation of the results.
  - Making decisions on the basis of such analysis

Role of statisticians

- To guide the design of an experiment or survey prior to data collection

- To analyze data using proper statistical procedures and techniques

- To present and interpret the results to researchers and other decision makers



Sources of data

Records

Surveys

Experiments

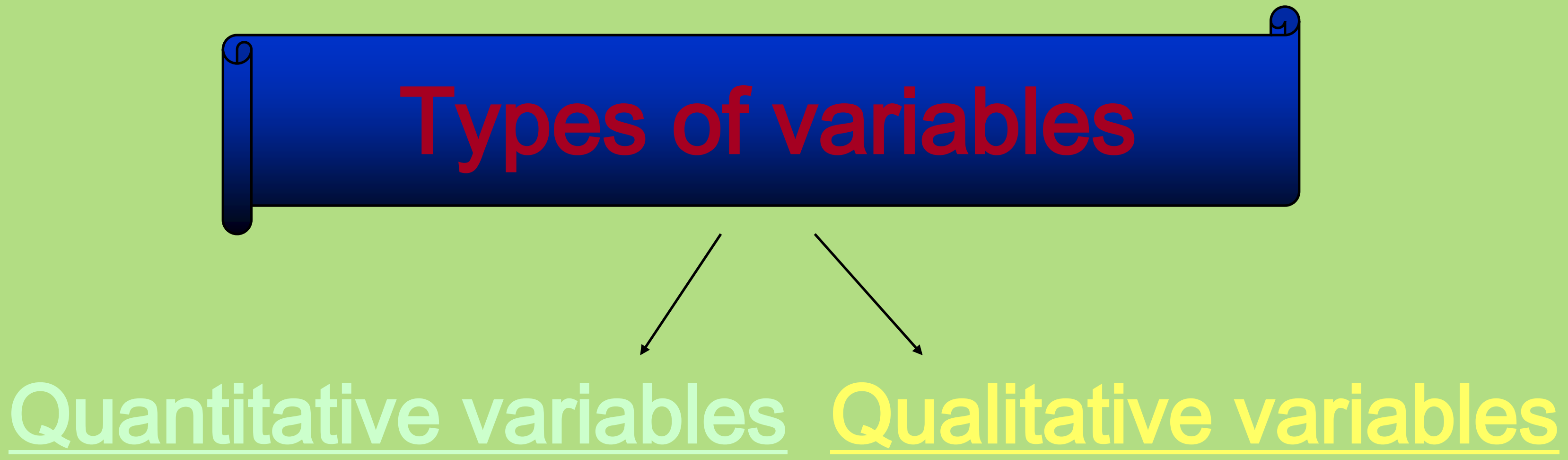
Comprehensive

Sample

Types of data

Constan
t

Variables



Types of variables

Quantitative variables

Qualitative variables

Methods of presentation of data

➊ Numerical presentation

➋ Graphical presentation

➌ Mathematical presentation

1- Numerical presentation

Tabular presentation (simple – complex)

Simple frequency distribution Table (S.F.D.T.)

Title

Name of variable (Units of variable)	Frequency	%
- - Categories -		
Total		

Table (I): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their ABO blood groups

Blood group	Frequency	%
A	12	24
B	18	36
AB	5	10
O	15	30
Total	50	100

Table (II): Distribution of 50 patients at the surgical department of Alexandria hospital in May 2008 according to their age

Age (years)	Frequency	%
20-<30	12	24
30-	18	36
40-	5	10
50+	15	30
Total	50	100

Complex frequency distribution Table

Table (IV): Distribution of 60 patients at the chest department of Alexandria hospital in May 2008 according to smoking & lung cancer

Smoking	Lung cancer				Total	
	positive		negative			
	No.	%	No.	%	No.	%
Smoker	15	65.2	8	34.8	23	100
Non smoker	5	13.5	32	86.5	37	100
Total	20	33.3	40	66.7	60	100

2- Graphical presentation

❶ *Graphs drawn using coordinates*

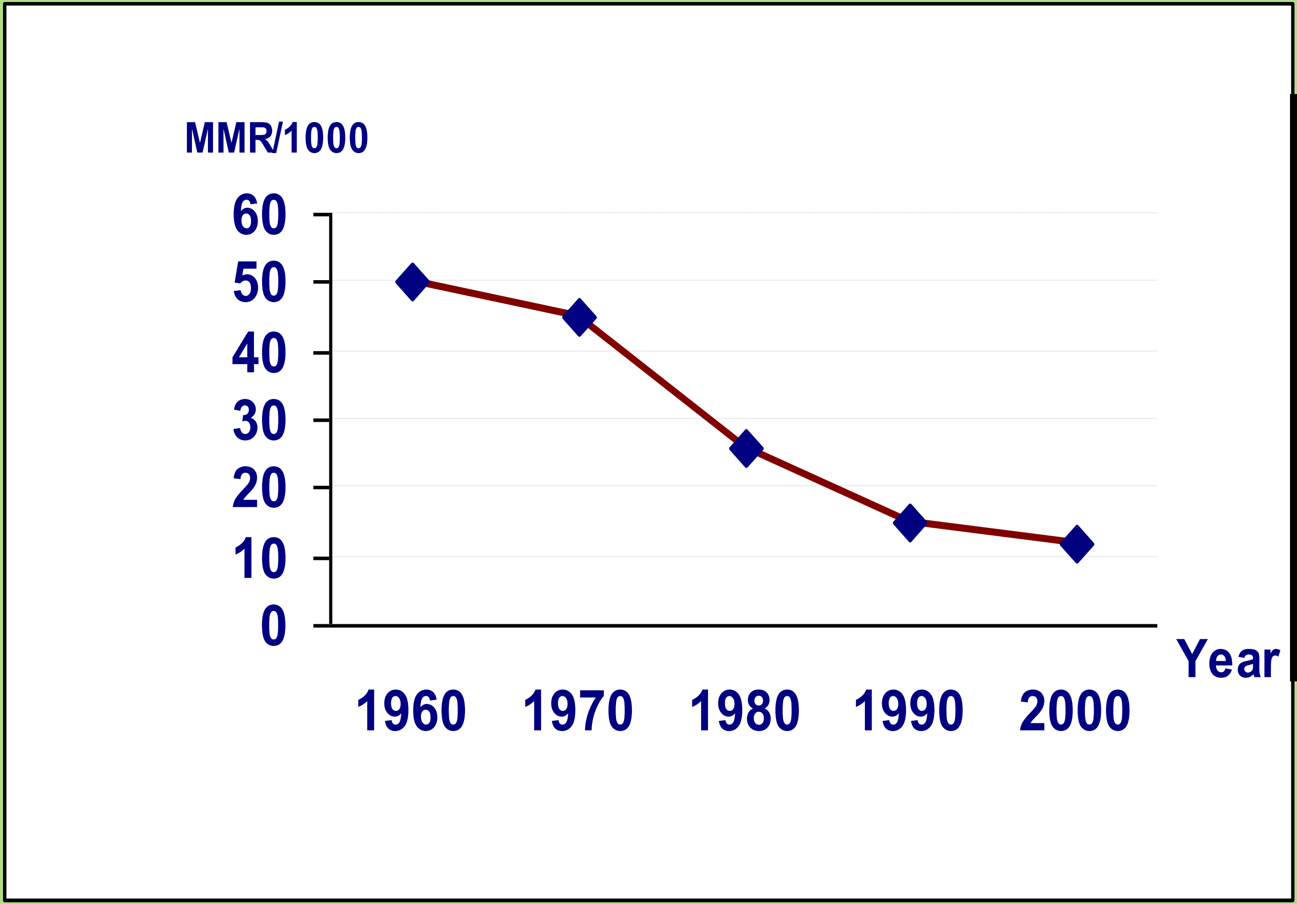
- Line graph
- Frequency polygon
- Frequency curve
- Histogram
- Bar graph
- Scatter plot

❷ *Pie chart*

❸ *Statistical maps*

rules

Line Graph



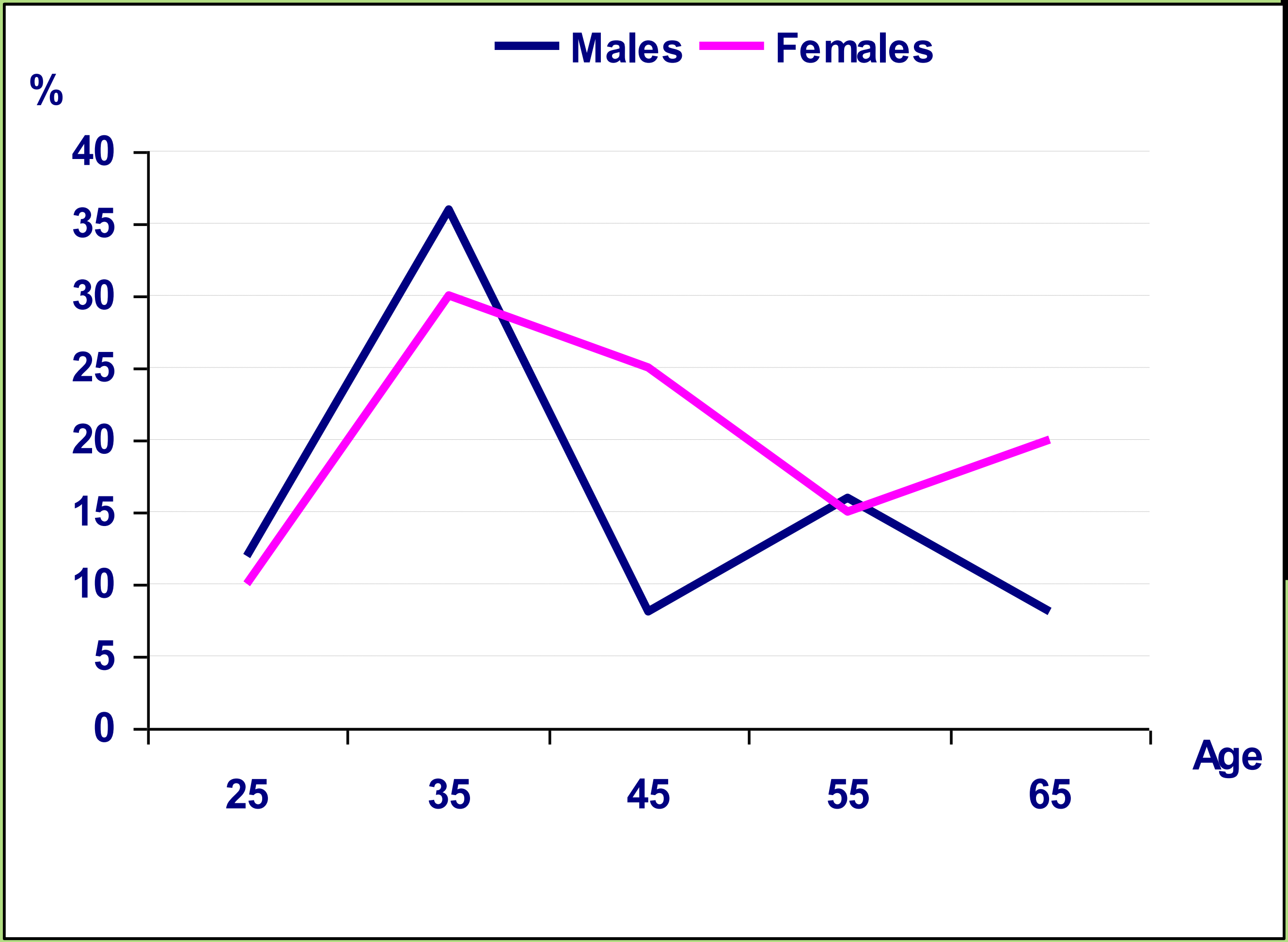
Year	MMR
1960	50
1970	45
1980	26
1990	15
2000	12

Figure (1): Maternal mortality rate of (country), 1960-2000

Frequency polygon

Age (years)	Sex		Mid-point of interval
	Males	Females	
20 - 30	3 (12%)	2 (10%)	(20+30) / 2 = 25
30 – 40	9 (36%)	6 (30%)	(30+40) / 2 = 35
40- 50	7 (8%)	5 (25%)	(40+50) / 2 = 45
50 - 60	4 (16%)	3 (15%)	(50+60) / 2 = 55
60 - 70	2 (8%)	4 (20%)	(60+70) / 2 = 65
Total	25(100%)	20(100%)	

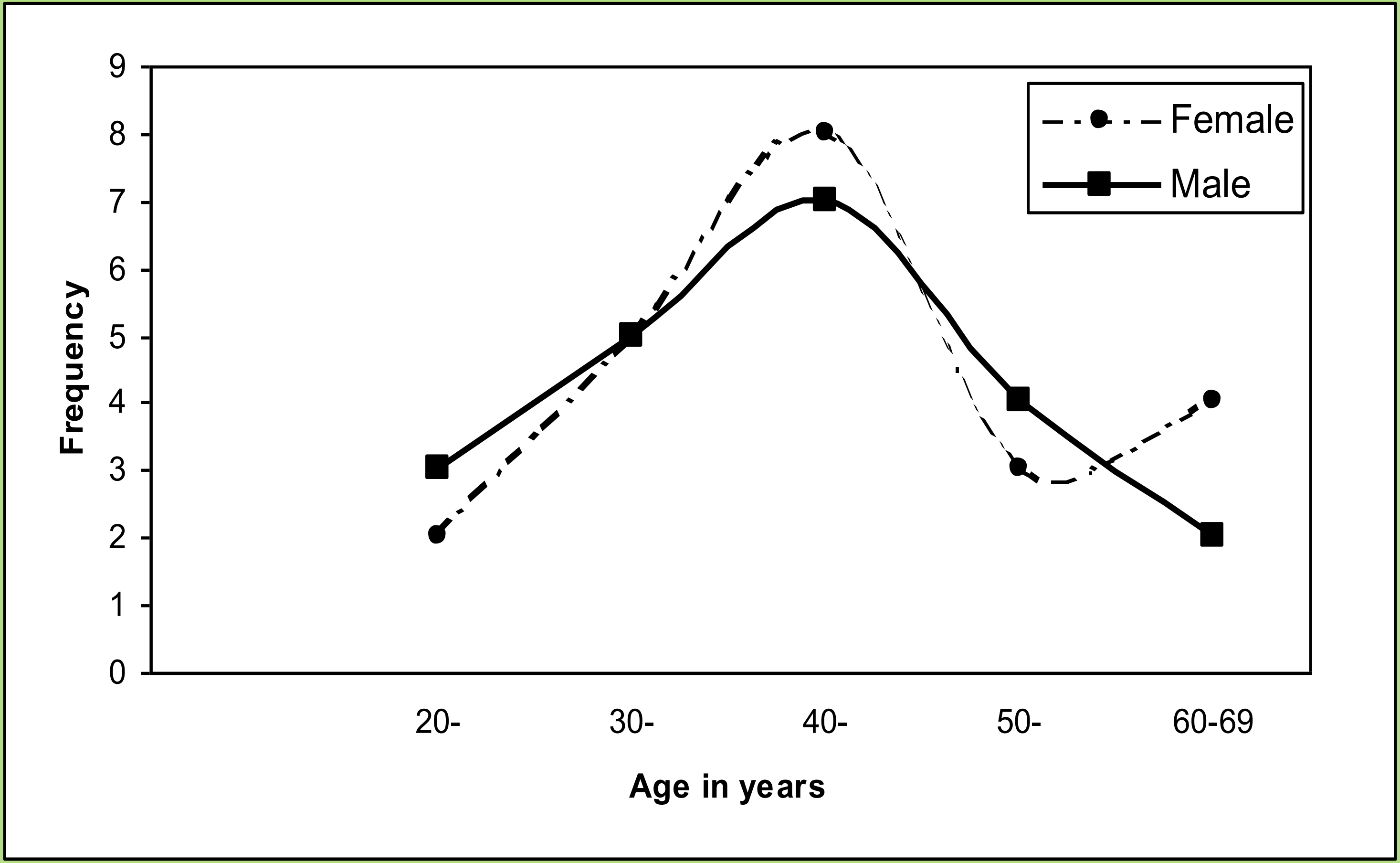
Frequency polygon



Age	Sex		M-P
	M	F	
20-	(12%)	(10%)	25
30-	(36%)	(30%)	35
40-	(8%)	(25%)	45
50-	(16%)	(15%)	55
60-70	(8%)	(20%)	65

Figure (2): Distribution of 45 patients at (place) , in (time) by age and sex

Frequency curve



Histogram

Distribution of a group of cholera patients by age

Age (years)	Frequency	%
25-	3	14.3
30-	5	23.8
40-	7	33.3
45-	4	19.0
60-65	2	9.5
Total	21	100

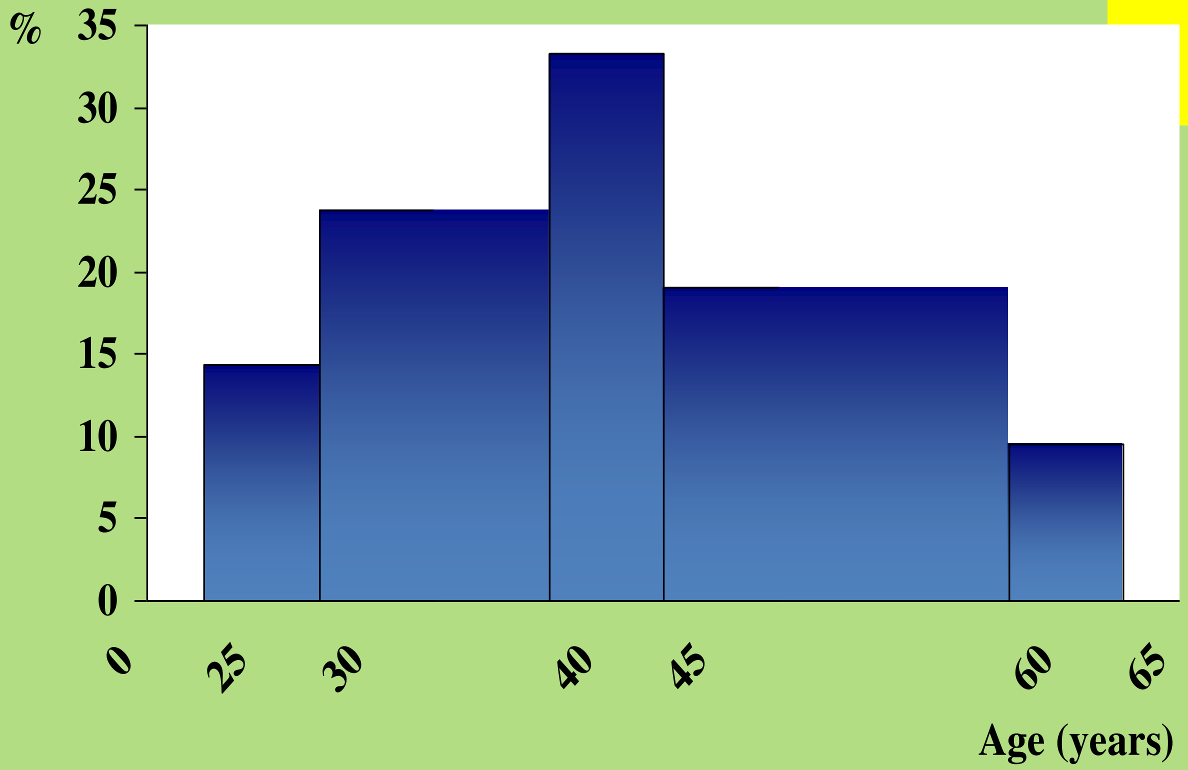
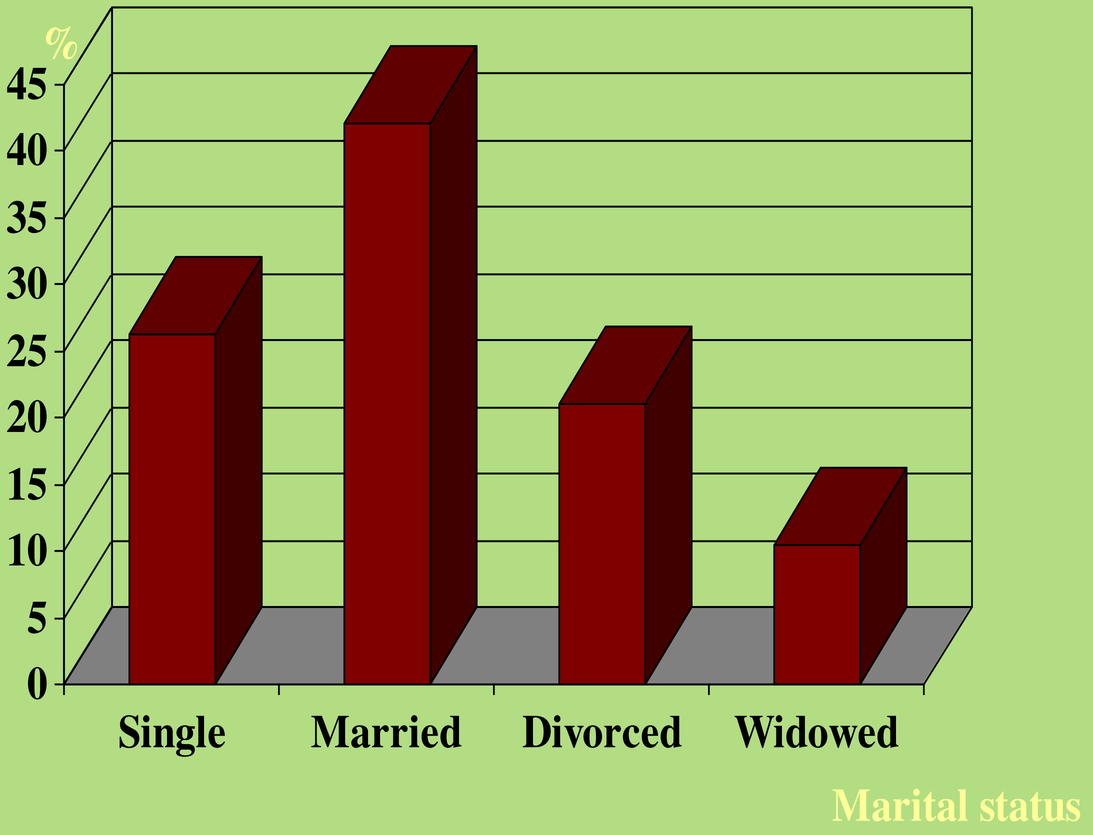
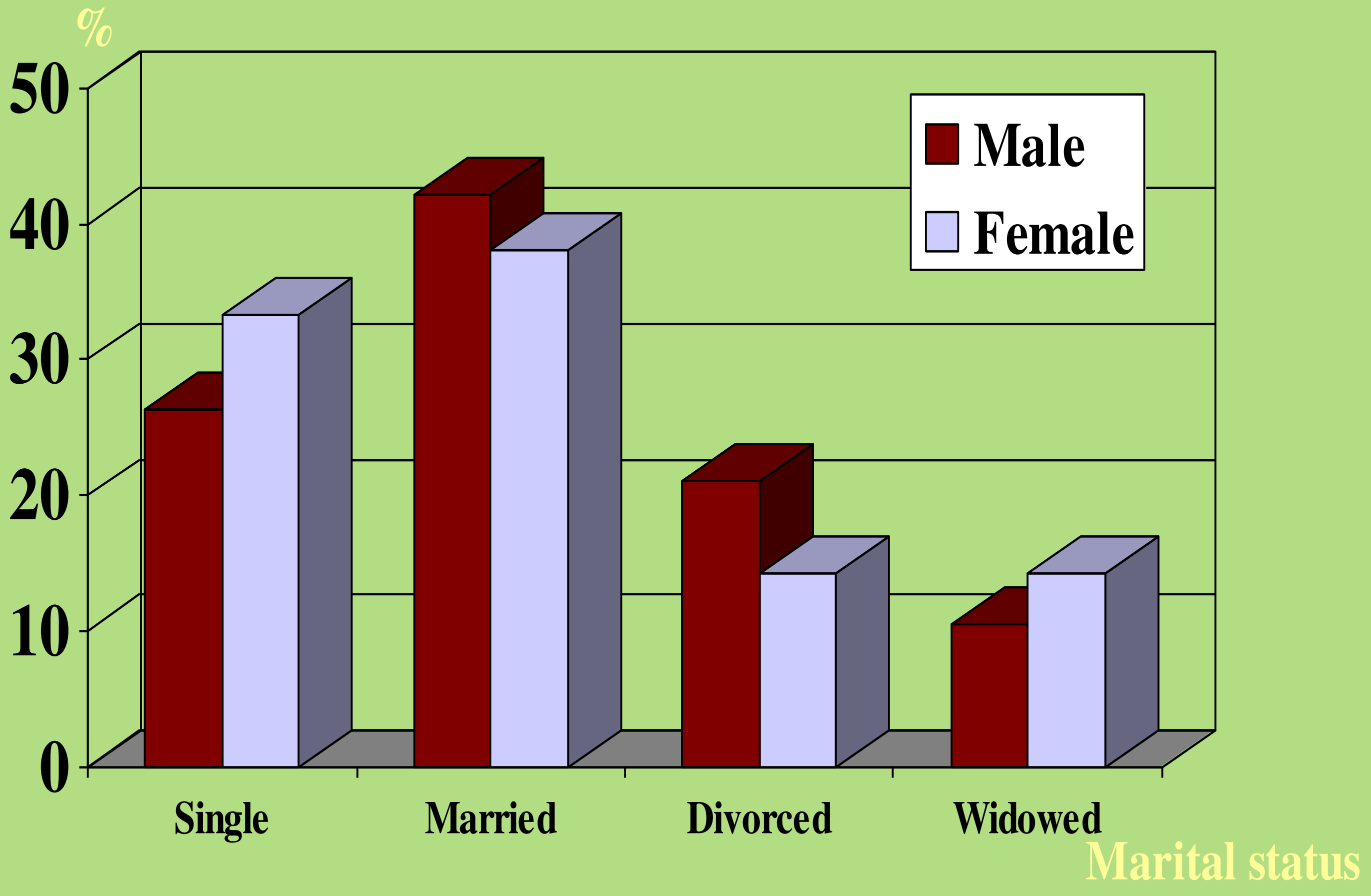


Figure (2): Distribution of 100 cholera patients at (place) , in (time) by age

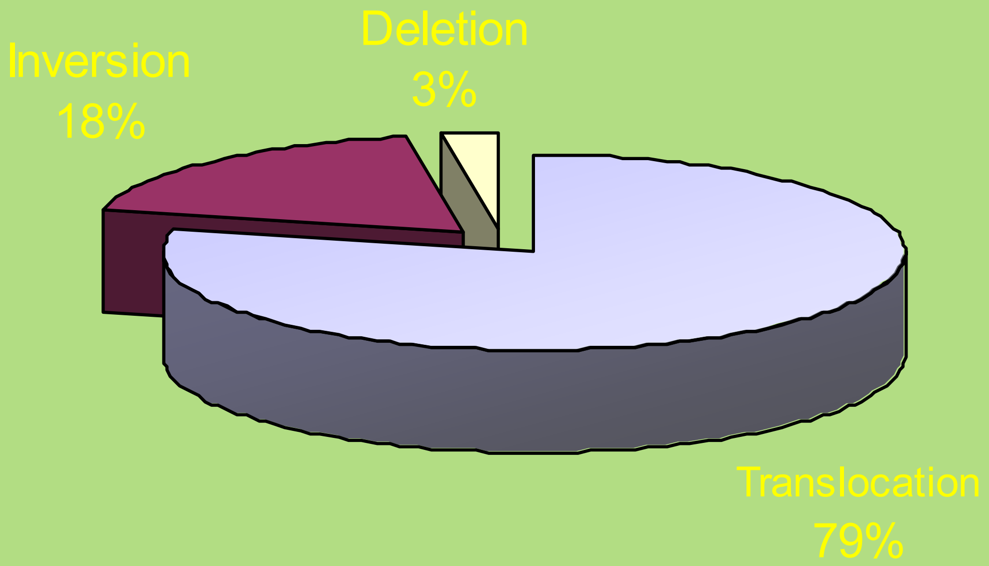
Bar chart


%
45
40
35
30
25
20
15
10
5
0
Single
Married
Divorced
Widowed
Marital status

Bar chart


%
50
40
30
20
10
0
Male
Female
Single
Married
Divorced
Widowed
Marital status

Pie chart



3-Mathematical presentation

Summery statistics

- Measures of location

  1- Measures of central tendency

  2- Measures of non central locations

  (Quartiles, Percentiles)

- Measures of dispersion

Descriptive measures

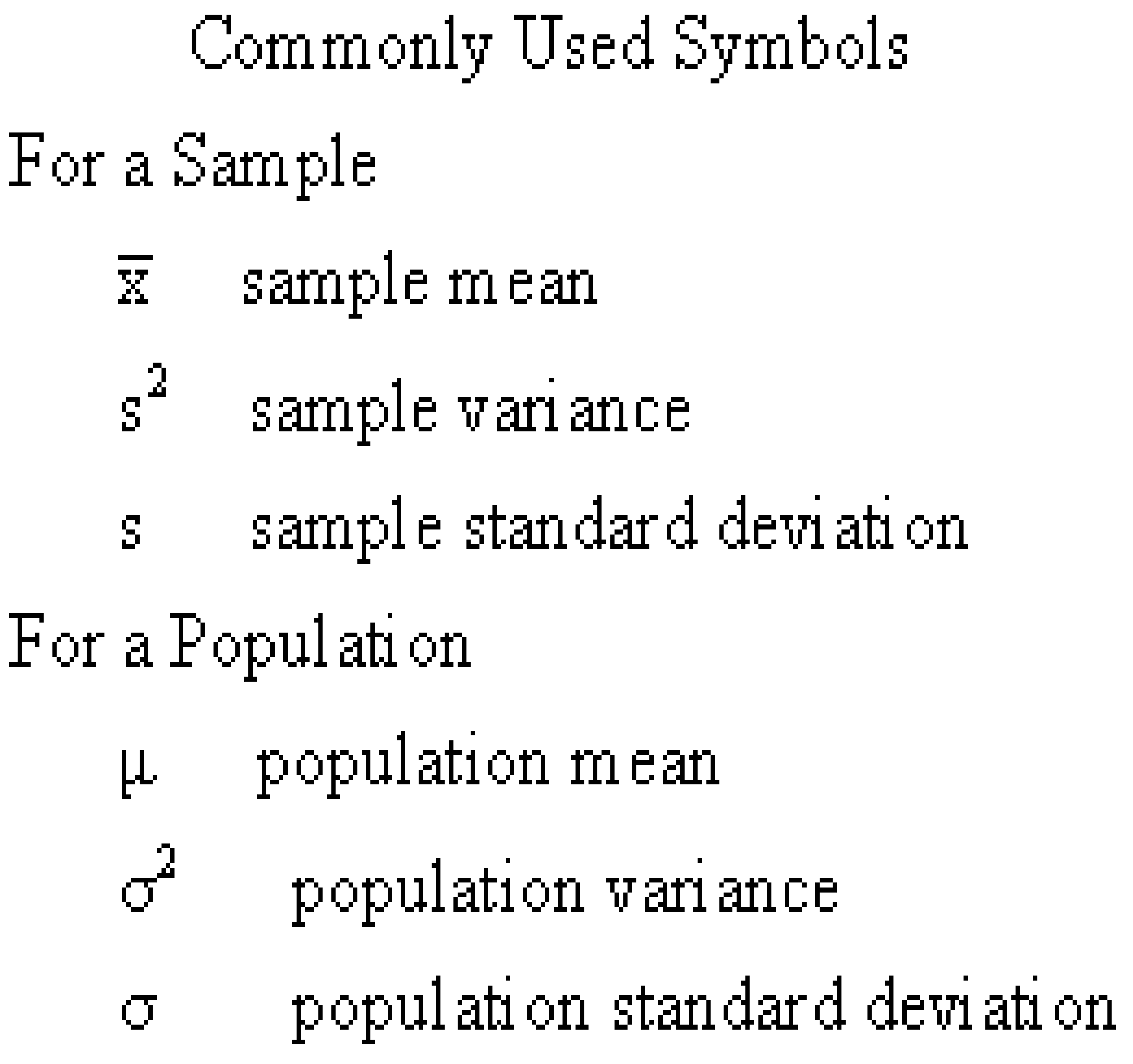
- A ***descriptive measure*** is a single number that is used to describe a set of data.
- Descriptive measures include ***measures of central tendency*** and ***measures of dispersion***.

Measures of Central Tendency

- ***Central tendency*** is a property of the data that they tend to be clustered about a center point.
- ***Measures of central tendency*** include:
  - **mean** (generally not part of the data set)
  - **median** (may be part of the data set)
  - **mode** (always part of the data set)

Measures of Dispersion

- ***Dispersion*** is a property of the data that they tend to be spread out.
- ***Measures of dispersion*** include:
  - **range**
  - **variance**
  - **standard deviation**



Commonly Used Symbols

For a Sample

$\bar{x}$	sample mean
s^2	sample variance
s	sample standard deviation

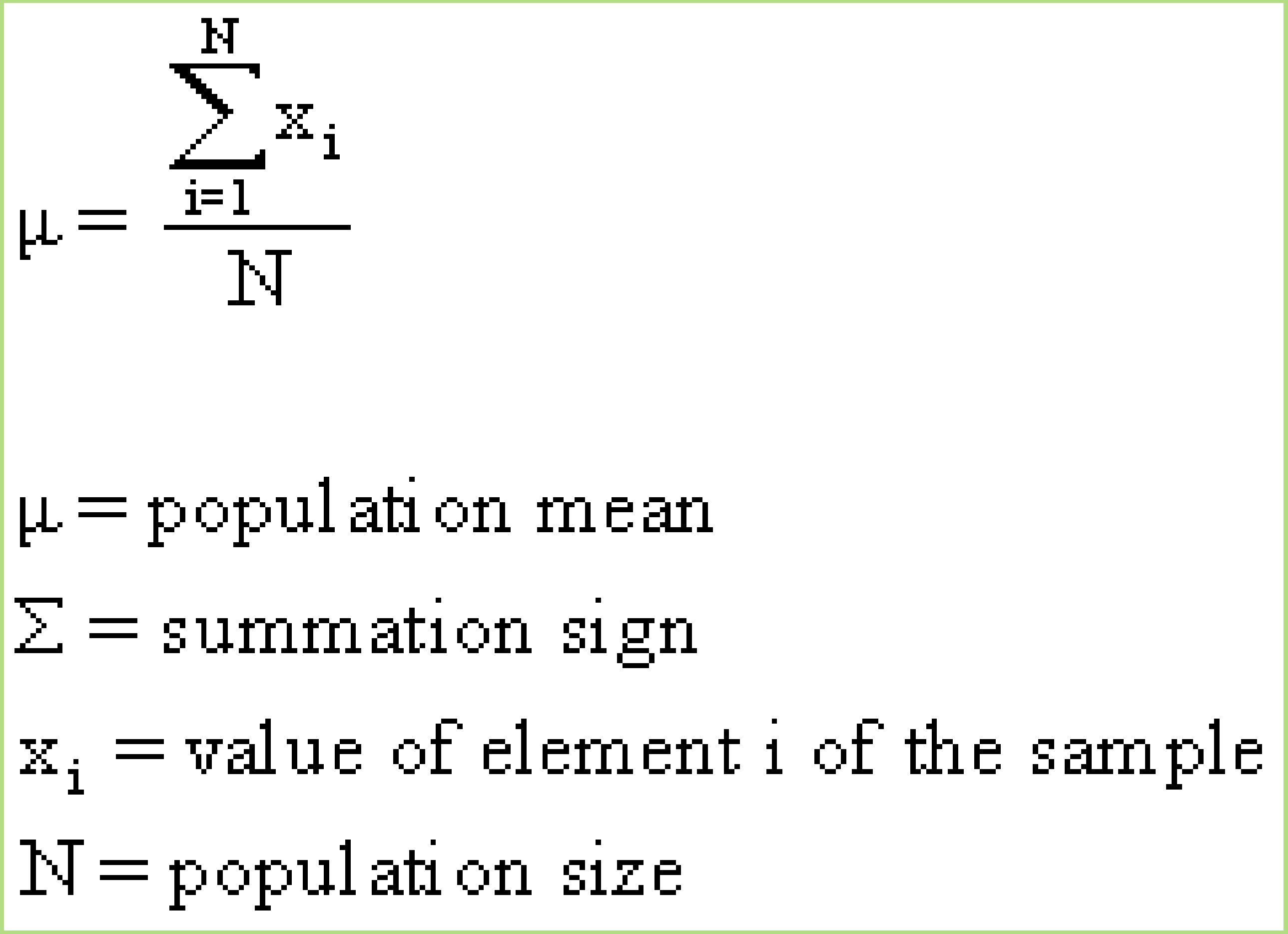
For a Population

μ	population mean
σ^2	population variance
σ	population standard deviation

Arithmetic mean

- The ***mean*** or ***arithmetic mean*** is the "average" which is obtained by adding all the values in a sample or population and dividing them by the number of values.

General formula--population mean



$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

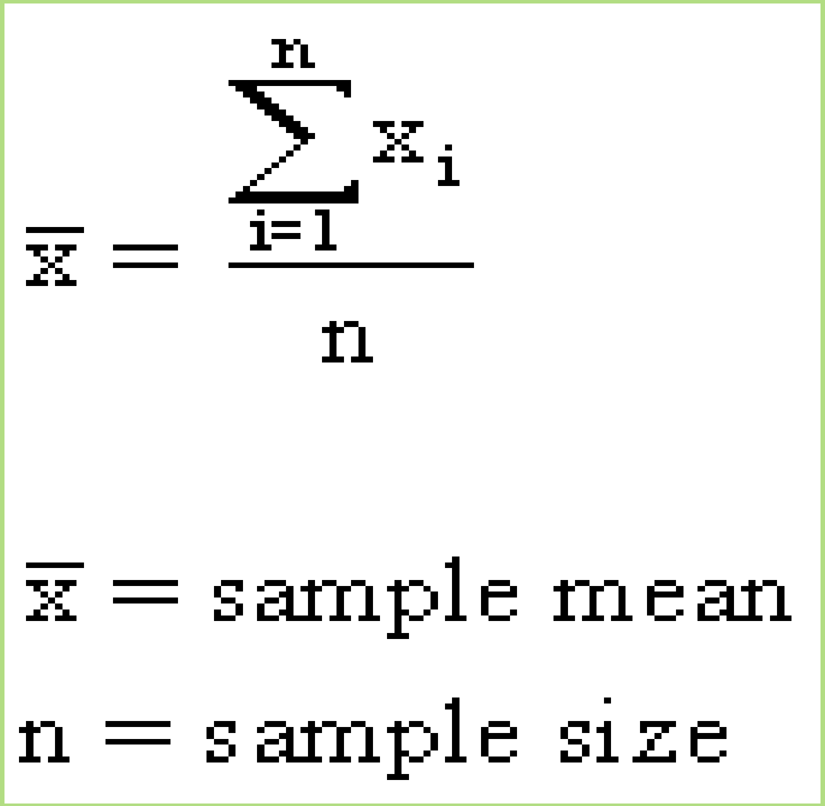
μ = population mean

Σ = summation sign

x_i = value of element i of the sample

N = population size

General formula--sample mean



$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$\overline{x}$ = sample mean

n = sample size

Properties of the mean

1. ***Uniqueness*** -- For a given set of data there is one and only one mean.
2. ***Simplicity*** -- The mean is easy to calculate.
3. ***Affected by extreme values*** -- The mean is influenced by each value. Therefore, extreme values can distort the mean.

Median

- The ***median*** is the value that divides the set of data into two equal parts. It is the midpoint of the data set.
- The number of values equal to or greater than the median equals the number of values less than or equal to the median.

Finding the median

1. Arrange (sort) the data in order of increasing value in a sorted list.

2. Find the median.

a. Odd number of values (n is odd)

$$\text{median} = \frac{n+1}{2}$$

Finding the median

b. Even number of values
(n is even)

median = average of the two
values in the middle

Properties of the median

1. ***Uniqueness*** -- There is only one median for each set of data.
2. ***Simplicity*** -- It is easy to calculate.
3. ***Effect of extreme values*** -- The median is not as drastically affected by extreme values as is the mean.

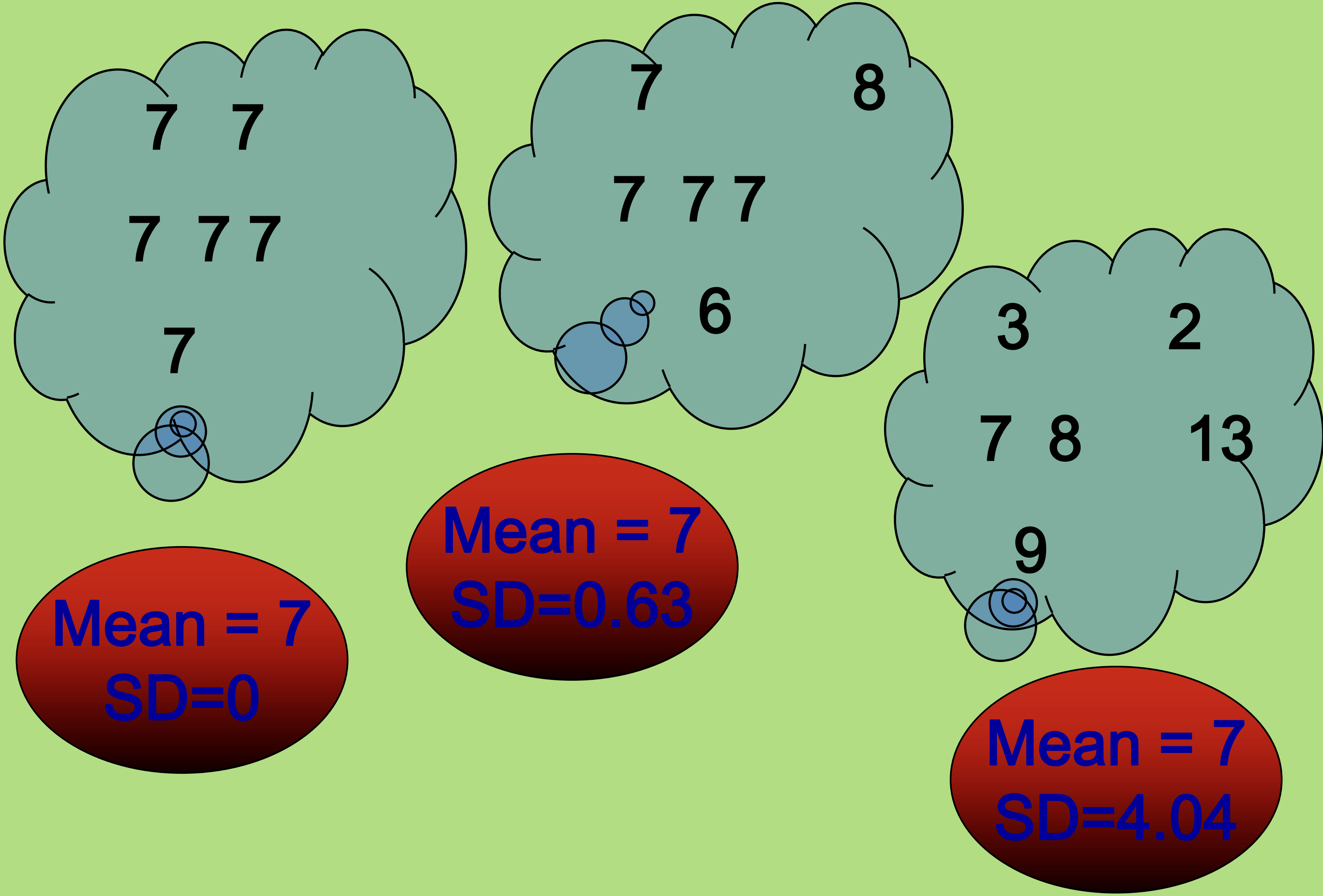
Mode

- The ***mode*** is the value that occurs most often in a set of data.
- It is possible to have more than one mode or no mode.

Variability of data

- ***Dispersion*** refers to the variety exhibited by the values of the data. The amount may be small when the values are close together.

Standard deviation SD



7 7
7 7 7
7

Mean = 7
SD=0

7 8
7 7 7
6

Mean = 7
SD=0.63

3 2
7 8 13
9

Mean = 7
SD=4.04

Range

- The ***range*** is the difference between the largest and smallest values in the set of observations.
- These values are often called the ***maximum*** and the ***minimum***.

Variance

- **Variance** is used to measure the dispersion of values relative to the mean.
- When values are close to their mean (narrow range) the dispersion is less than when there is scattering over a wide range.

Calculation of the sample variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

s^2 = sample variance

x_i = individual value

$\overline{x}$ = sample mean

n = number of values

Variance of a population

$$\sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

= population variance

σ^2 = population size

μ = population mean

Degrees of freedom

- In computing the variance there are n - 1 ***degrees of freedom*** because if n -1 values are known, the nth one is determined automatically.
- This is because all of the values of

$(x_i - \bar{x})$ must add to zero.

Differences in calculations

Values of s^2 and σ^2 are different

because s^2 divides by n-1

whereas σ^2 divides by N.

Sample standard deviation

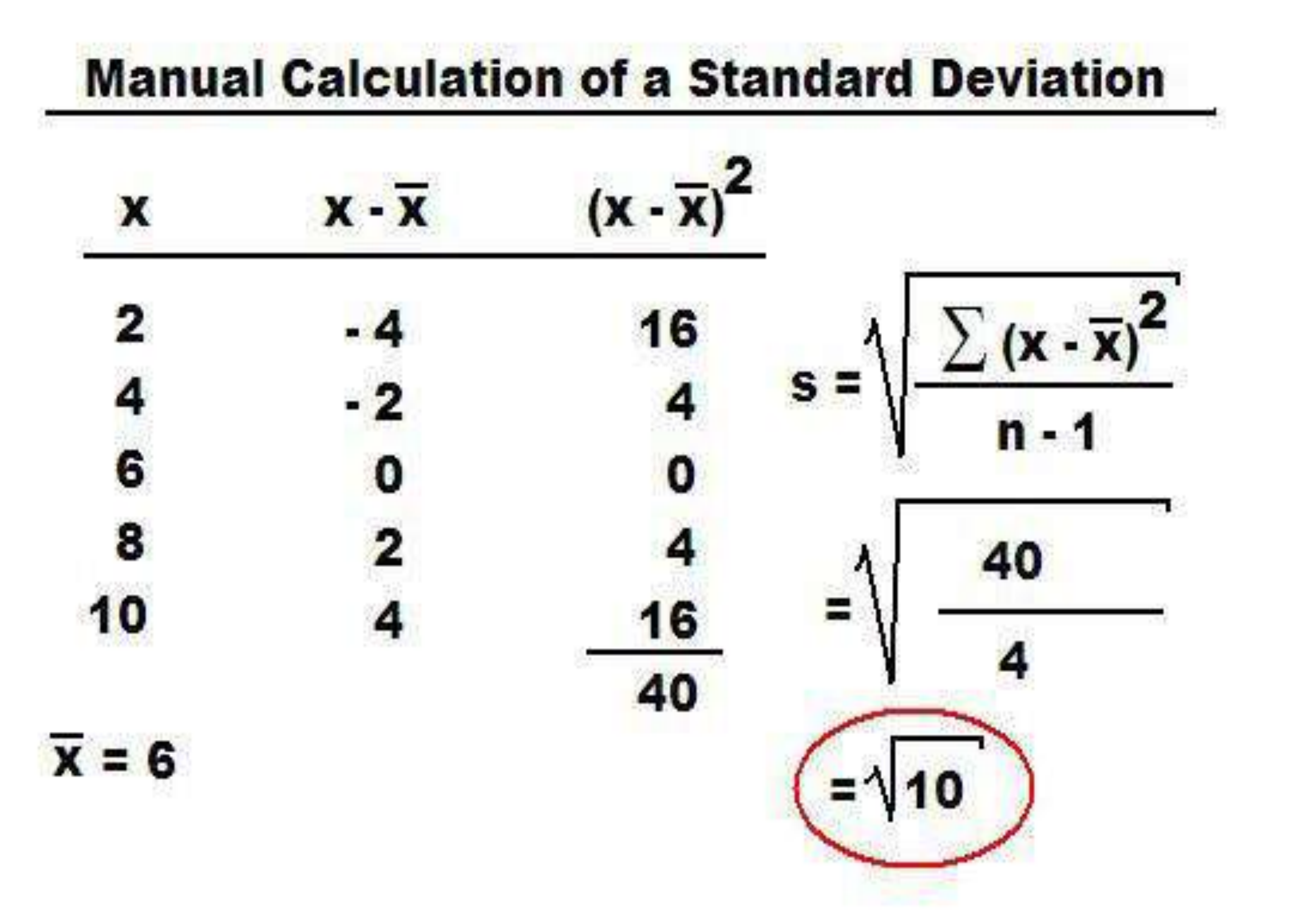
The ***standard deviation*** is the square root of the variance. The standard deviation expresses the dispersion in terms of the original units. Since the variance of a sas^2ple is , we take the square root.

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Population Standard Deviation

For a population, the standard deviation is σ which is the square root of the population variance.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$



Manual Calculation of a Standard Deviation

x	$x - \bar{x}$	$(x - \bar{x})^2$
2	-4	16
4	-2	4
6	0	0
8	2	4
10	4	16
		40

$\bar{x} = 6$

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

$$= \sqrt{\frac{40}{4}}$$

$$= \sqrt{10}$$

Coefficient of variation

Coefficient of variation is a measure of the relative amount of variation as opposed to the absolute variation.

$$\mathrm{C.V.} = \frac{s}{\bar{x}}(100)$$

C.V. is independent of the units of measure. It can be useful for comparing different results from people investigating the same variable.