Description of Data (Summary and Variability measures)

Dr.Shaik Shaffi Ahamed Ph.d.,
Associate Professor
Deparment of Family & Community Medicine
College of Medicine, KSU

Objectives of this session

- Able to understand how to summarize the data.
- Able to understand how to measure the variability of the data.
- Able to use and interpret appropriately the different summary and variability measures.

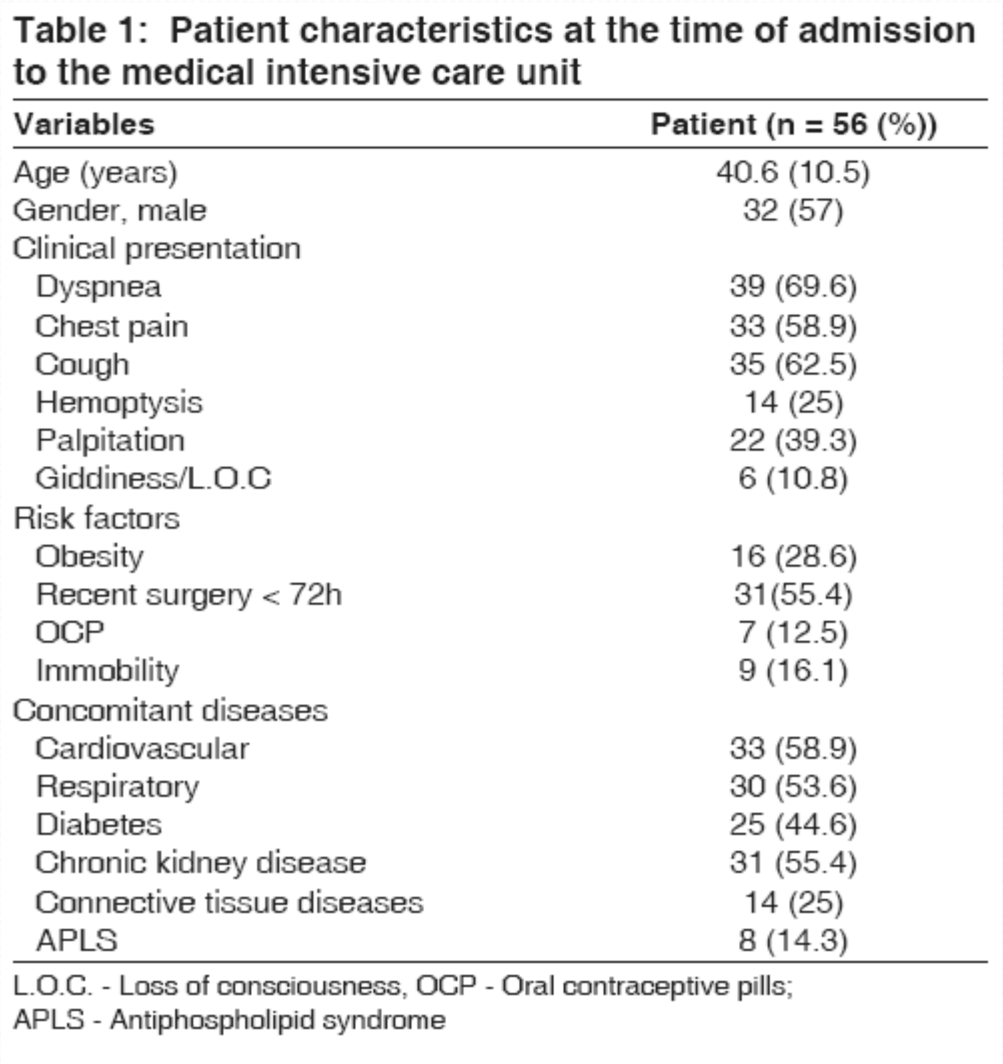


Table 1: Patient characteristics at the time of admission to the medical intensive care unit

Variables	Patient (n = 56 (%))
Age (years)	40.6 (10.5)
Gender, male	32 (57)
Clinical presentation	
Dyspnea	39 (69.6)
Chest pain	33 (58.9)
Cough	35 (62.5)
Hemoptysis	14 (25)
Palpitation	22 (39.3)
Giddiness/L.O.C	6 (10.8)
Risk factors	
Obesity	16 (28.6)
Recent surgery < 72h	31(55.4)
OCP	7 (12.5)
Immobility	9 (16.1)
Concomitant diseases	
Cardiovascular	33 (58.9)
Respiratory	30 (53.6)
Diabetes	25 (44.6)
Chronic kidney disease	31 (55.4)
Connective tissue diseases	14 (25)
APLS	8 (14.3)

L.O.C. - Loss of consciousness, OCP - Oral contraceptive pills;
APLS - Antiphospholipid syndrome

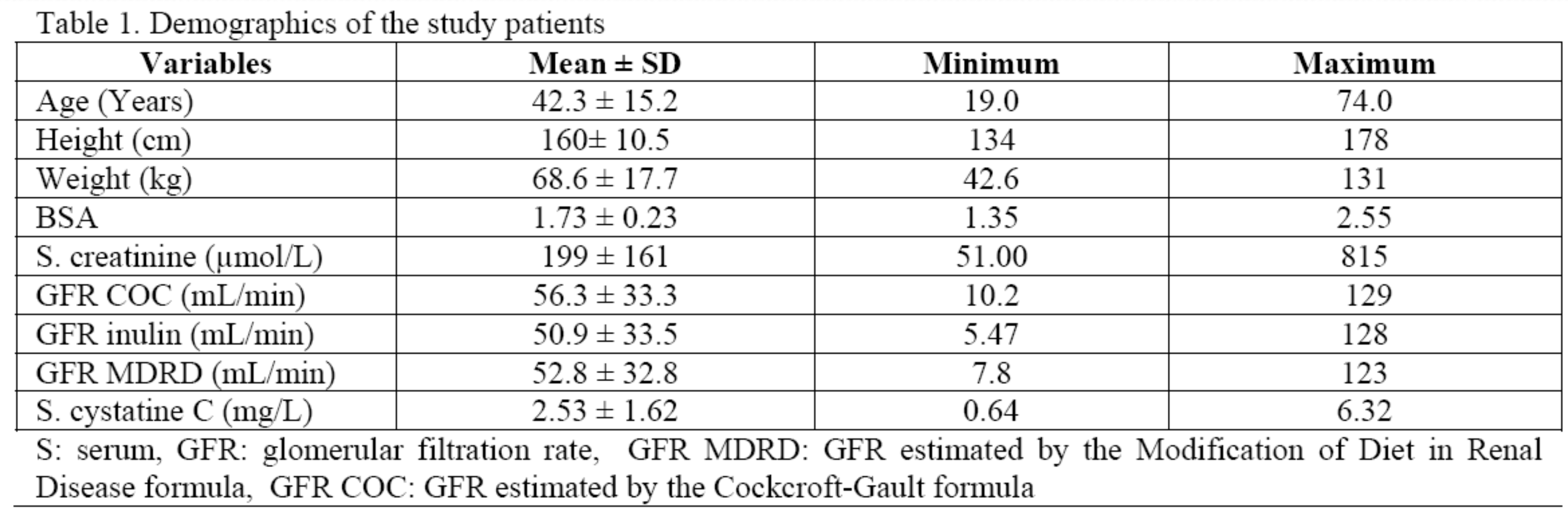
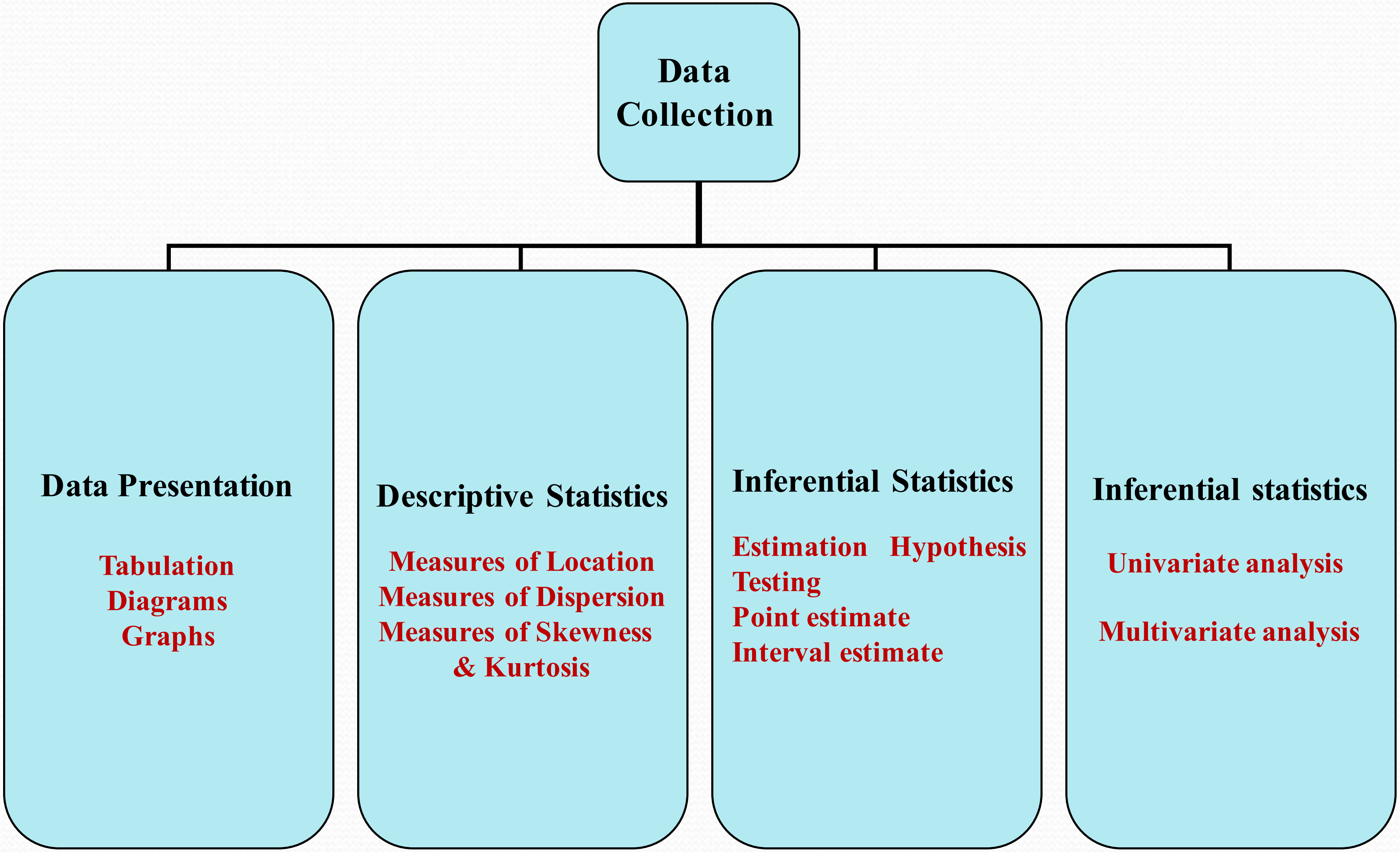


Table 1. Demographics of the study patients

Variables	Mean ± SD	Minimum	Maximum
Age (Years)	42.3 ± 15.2	19.0	74.0
Height (cm)	160± 10.5	134	178
Weight (kg)	68.6 ± 17.7	42.6	131
BSA	1.73 ± 0.23	1.35	2.55
S. creatinine (µmol/L)	199 ± 161	51.00	815
GFR COC (mL/min)	56.3 ± 33.3	10.2	129
GFR inulin (mL/min)	50.9 ± 33.5	5.47	128
GFR MDRD (mL/min)	52.8 ± 32.8	7.8	123
S. cystatine C (mg/L)	2.53 ± 1.62	0.64	6.32

S: serum, GFR: glomerular filtration rate, GFR MDRD: GFR estimated by the Modification of Diet in Renal Disease formula, GFR COC: GFR estimated by the Cockcroft-Gault formula

Investigation



Data Collection

Data Presentation

Tabulation
Diagrams
Graphs

Descriptive Statistics

Measures of Location
Measures of Dispersion
Measures of Skewness & Kurtosis

Inferential Statistics

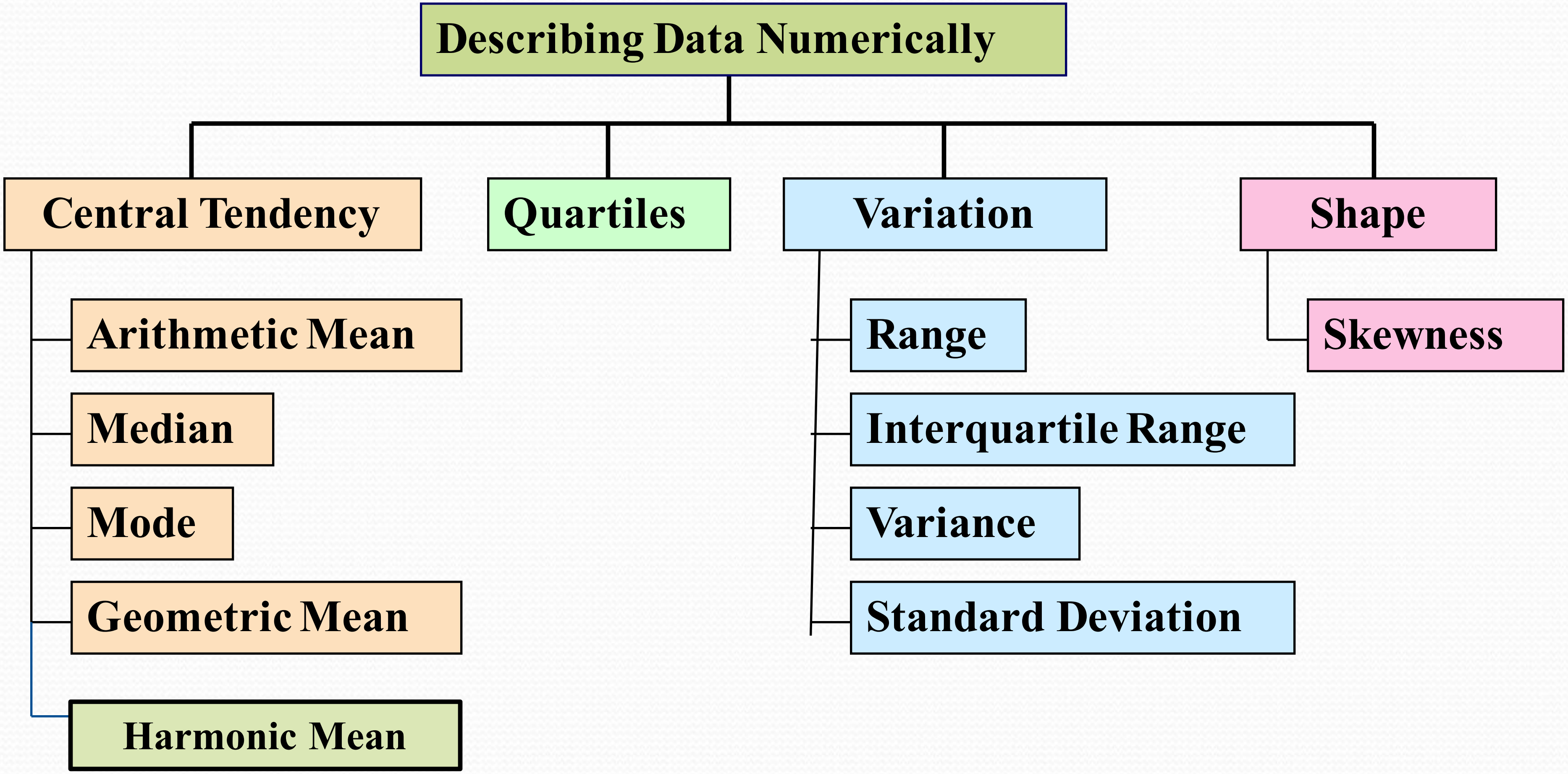
Estimation Hypothesis Testing
Point estimate
Interval estimate

Inferential statistics

Univariate analysis

Multivariate analysis

Summary & Variability Measures



Describing Data Numerically

- Central Tendency
  - Arithmetic Mean
  - Median
  - Mode
  - Geometric Mean
  - Harmonic Mean
- Quartiles
- Variation
  - Range
  - Interquartile Range
  - Variance
  - Standard Deviation
- Shape
  - Skewness

Measures of Central Tendency

- A statistical measure that identifies a single score as representative for an entire distribution. The goal of central tendency is to find the single score that is most typical or most representative of the entire group
- **There are three common measures of central tendency:**
  - **the mean**
  - **the median**
  - **the mode**

Calculating the Mean

- Calculate the mean of the following data:
  1 5 4 3 2

- Sum the scores (ΣX):
  $1 + 5 + 4 + 3 + 2 = 15$

- Divide the sum ($\Sigma X = 15$) by the number of scores ($N = 5$): $15 / 5 = 3$

- Mean = $\overline{X} = 3$

Simple Frequency Distributions

raw-score distribution

Name	X
Student1	20
Student2	23
Student3	15
Student4	21
Student5	15
Student6	21
Student7	15
Student8	20

→

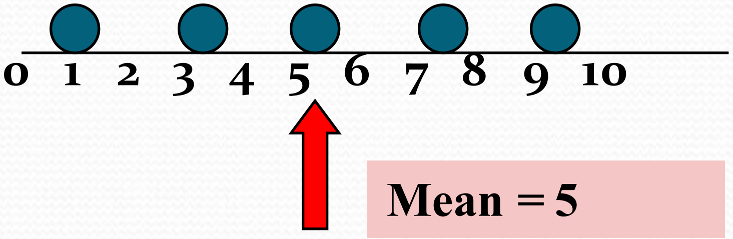
frequency distribution

f	X
3	15
2	20
2	21
1	23

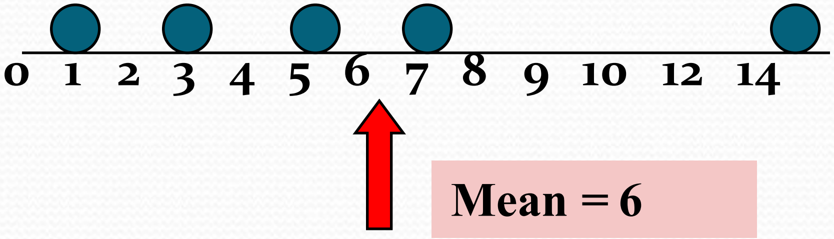
Mean $$\overline{\chi} = \frac{\Sigma f\chi}{N}$$

Mean (Arithmetic Mean) *(continued)*

- The most common measure of central tendency
- Affected by extreme values (outliers)



Mean = 5



Mean = 6

The Median

- The *median* is simply another name for the 50th percentile

- It is the score in the middle; half of the scores are larger than the median and half of the scores are smaller than the median

How To Calculate the Median

- Conceptually, it is easy to calculate the median

- Sort the data from highest to lowest

- Find the score in the middle
  - middle = (N + 1) / 2
  - If N, the number of scores is even, the median is the average of the middle two scores

Median Example

- What is the median of the following scores:
  24 18 19 42 16 12

- Sort the scores:
  42 24 19 18 16 12

- Determine the middle score:
  middle = (N + 1) / 2 = (6 + 1) / 2 = 3.5

- Median = average of 3rd and 4th scores:
  (19 + 18) / 2 = 18.5

Median Example

- What is the median of the following scores:
  10 8 14 15 7 3 3 8 12 10 9

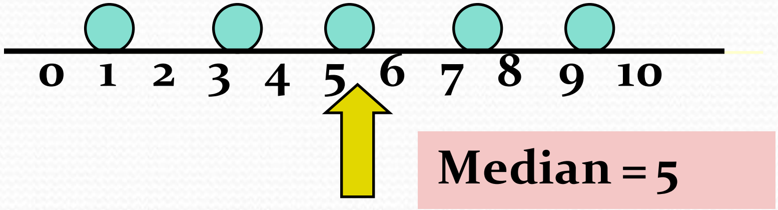
- Sort the scores:
  15 14 12 10 10 9 8 8 7 3 3

- Determine the middle score:
  middle = (N + 1) / 2 = (11 + 1) / 2 = 6

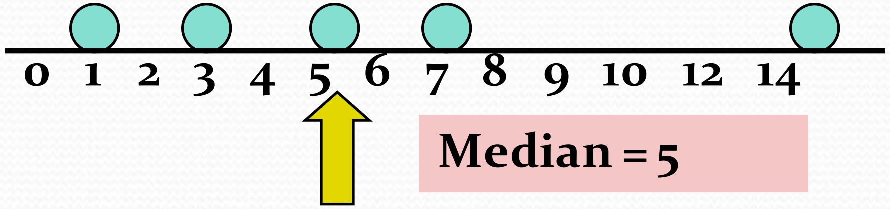
- Middle score = median = 9

Median

- Not affected by extreme values



Median = 5



Median = 5

- In an ordered array, the median is the “middle” number
  - If n or N is odd, the median is the middle number
  - If n or N is even, the median is the average of the two middle numbers

Measures of Central Tendency

Mean … the most frequently used but is sensitive to extreme scores

e.g. 1 2 3 4 5 6 7 8 9 10

Mean = 5.5 (median = 5.5)

e.g. 1 2 3 4 5 6 7 8 9 20

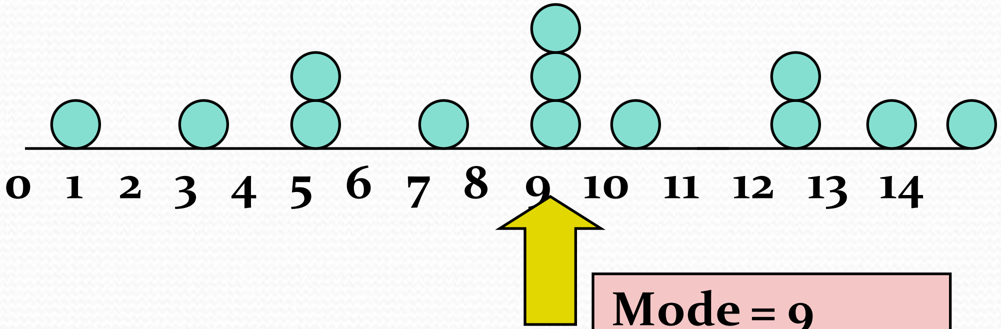
Mean = 6.5 (median = 5.5)

e.g. 1 2 3 4 5 6 7 8 9 100

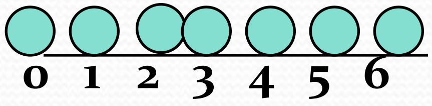
Mean = 14.5 (median = 5.5)

Mode

- Value that occurs most often
- Not affected by extreme values
- Used for either numerical or categorical(nominal)data
- There may be no mode
- There may be several modes



Mode = 9



No Mode

The Shape of Distributions

- **Distributions can be either <u>symmetrical</u> or <u>skewed</u>, depending on whether there are more frequencies at one end of the distribution than the other.**

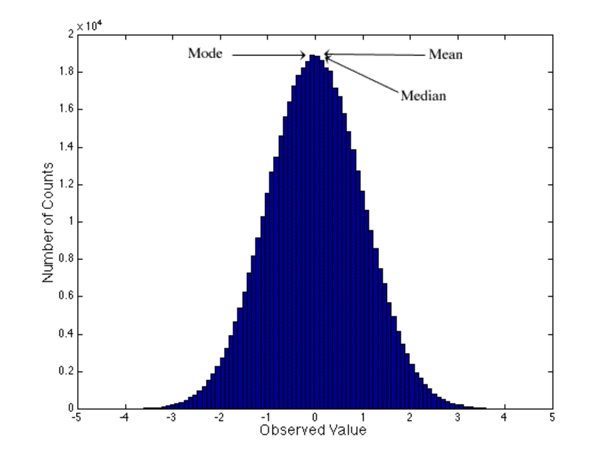
Symmetrical Distributions

- A distribution is symmetrical if the frequencies at the right and left tails of the distribution are identical, so that if it is divided into two halves, each will be the mirror image of the other.

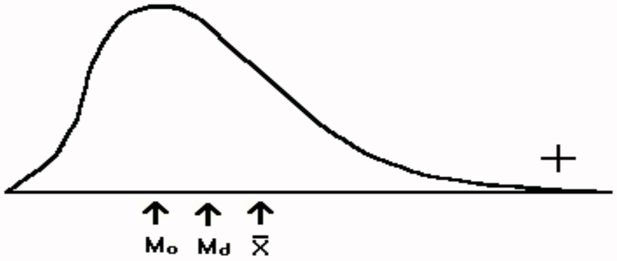
- In a symmetrical distribution the mean, median, and mode are identical.

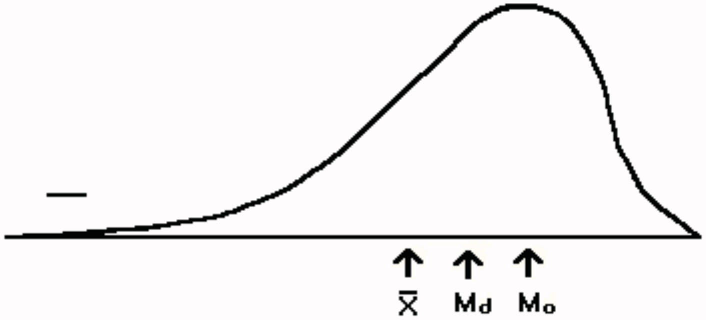
Distributions

- Bell-Shaped (also known as symmetric" or "normal")



- Skewed:
  - positively (skewed to the right) – it tails off toward larger values
  - negatively (skewed to the left) – it tails off toward smaller values





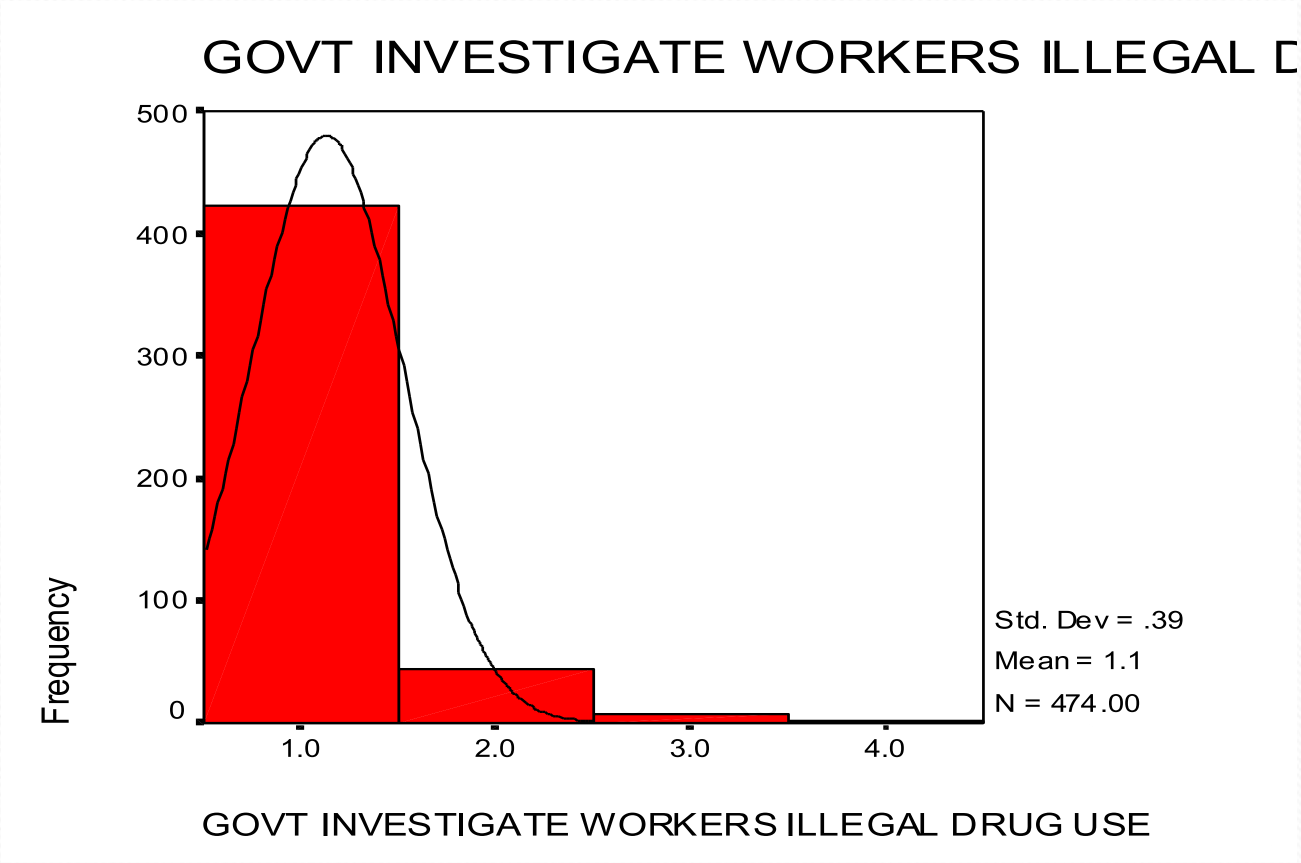
Skewed Distribution

Few extreme values on one side of the distribution or on the other.

- Positively skewed distributions: distributions which have few extremely high values (Mean>Median)

- Negatively skewed distributions: distributions which have few extremely low values(Mean<Median)

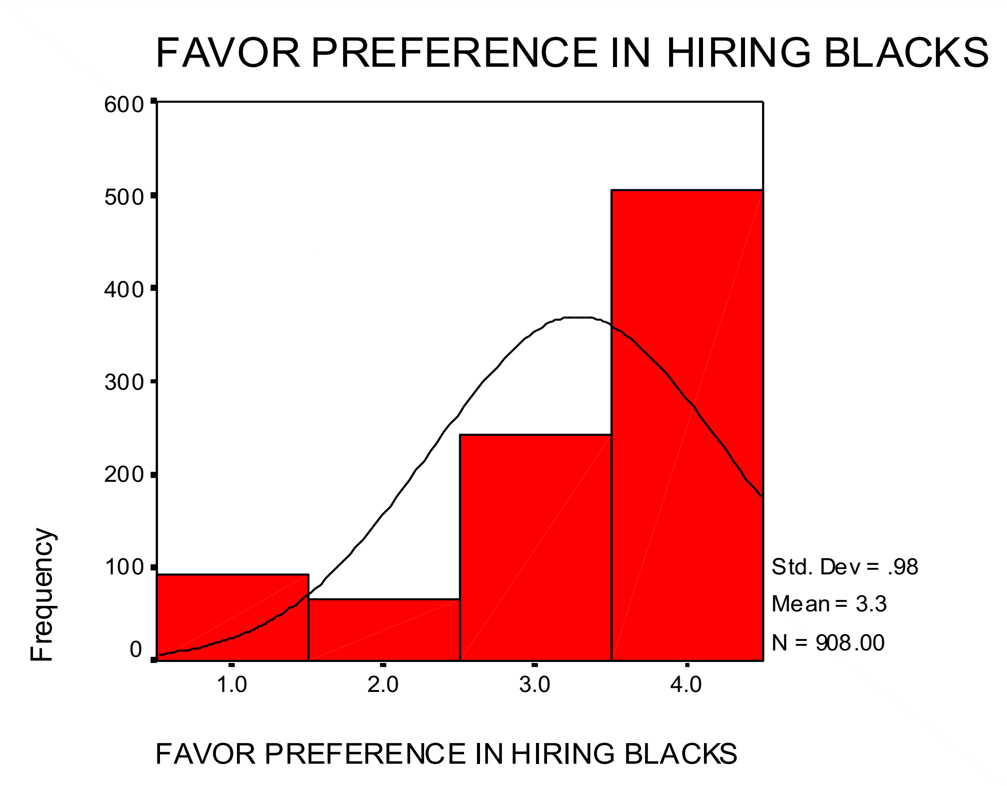
Positively Skewed Distribution



Mean =1.13

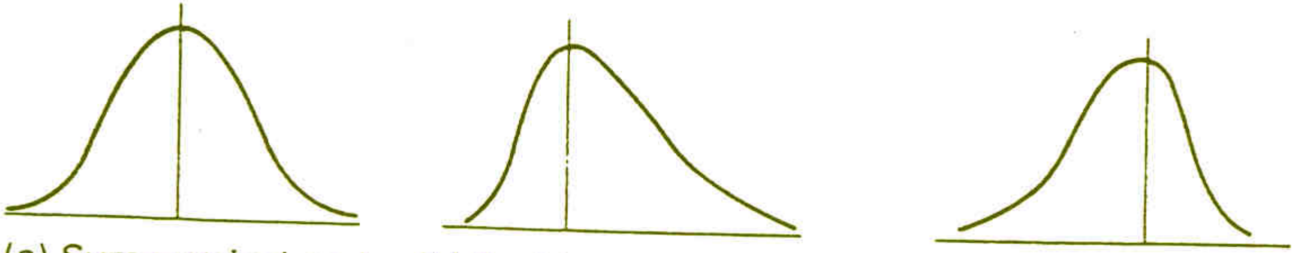
Median =1.0

Negatively Skewed distribution



Mean=3.3

Median=4.0

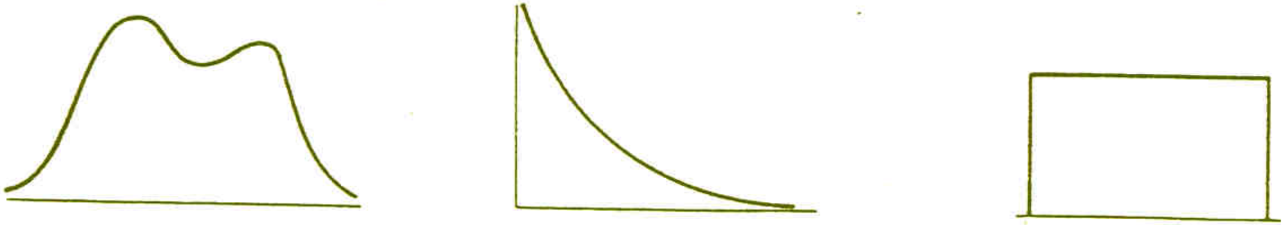


(a) Symmetrical and bell-shaped, e.g. height

(b) Positively skewed or skewed to the right, e.g. triceps skinfold measurement

(c) Negatively skewed or skewed to the left, e.g. period of gestation

Fig. 3.5 Three common shapes of frequency distributions with an example of each.



(a) Bimodal, e.g. hormone levels of males and females

(b) Reverse J-shaped, e.g. survival time after diagnosis of lung cancer

(c) Uniform, e.g. month of occurrence of disease with no seasonal pattern

e.g. Diabetes Mellitus

. 3.6 Three less-common shapes of frequency distributions with an example of each.

Choosing a Measure of Central tendency

- IF variable is Nominal..
- Mode
- IF variable is Ordinal...
- Mode or Median(or both)
- IF variable is Interval-Ratio and distribution is Symmetrical…
- Mode, Median or Mean
- IF variable is Interval-Ratio and distribution is Skewed…
- Mode or Median

EXAMPLE

(1) 7,8,9,10,11 n=5, $\sum x$=45, $\bar{x}$ =45/5=9

(2) 3,4,9,12,15 n=5,$\sum$ x=45, $\bar{x}$ =45/5=9

(3) 1,5,9,13,17 n=5,$\sum$ x=45, $\bar{x}$ =45/5=9

S.D. : (1) 1.58 (2) 4.74 (3) 6.32

Measures of Dispersion

Or

Measures of variability

Measures of Dispersion

Measures of dispersion summarize differences in the data, how the numbers differ from one another.

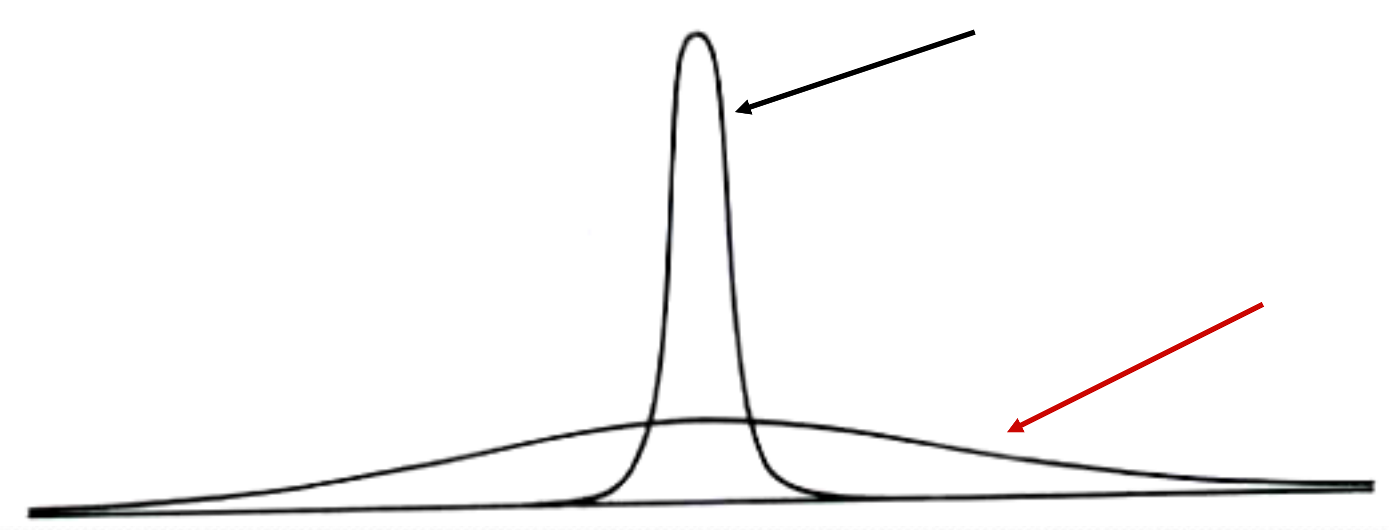
Series I: 70 70 70 70 70 70 70 70 70 70

Series II: 66 67 68 69 70 70 71 72 73 74

Series III: 1 19 50 60 70 80 90 100 110 120

Measures of Variability

- A single summary figure that describes the spread of observations within a distribution.



Measures of Variability

- Range
  - Difference between the smallest and largest observations.
- Interquartile Range
  - Range of the middle half of scores.
- Variance
  - Mean of all squared deviations from the mean.
- Standard Deviation
  - Rough measure of the average amount by which observations deviate from the mean. The square root of the variance.

Variability Example: Range

- Marks of students

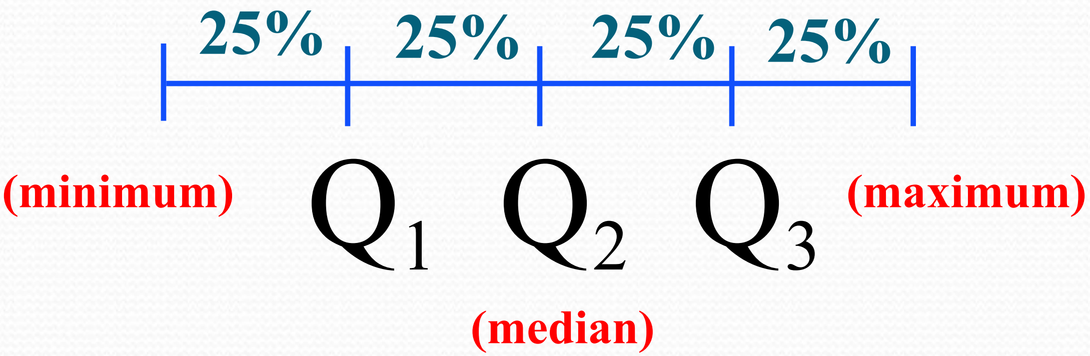
52, 76, 100, 36, 86, 96, 20, 15, 57, 64, 64, 80, 82, 83, 30, 31, 31, 31, 32, 37, 38, 38, 40, 40, 41, 42, 47, 48, 63, 63, 72, 79, 70, 71, 89

- Range: 100-15 = 85

Quartiles

Q_1, Q_2, Q_3

divides ranked scores into four equal parts



25% 25% 25% 25%

(minimum) Q_1 Q_2 Q_3 (maximum)

(median)

Quartiles: $$Q_1 = \frac{n+1}{4} th$$

$$Q_2 = \frac{2(n+1)}{4} = \frac{n+1}{2} th$$

$$Q_3 = \frac{3(n+1)}{4} th$$

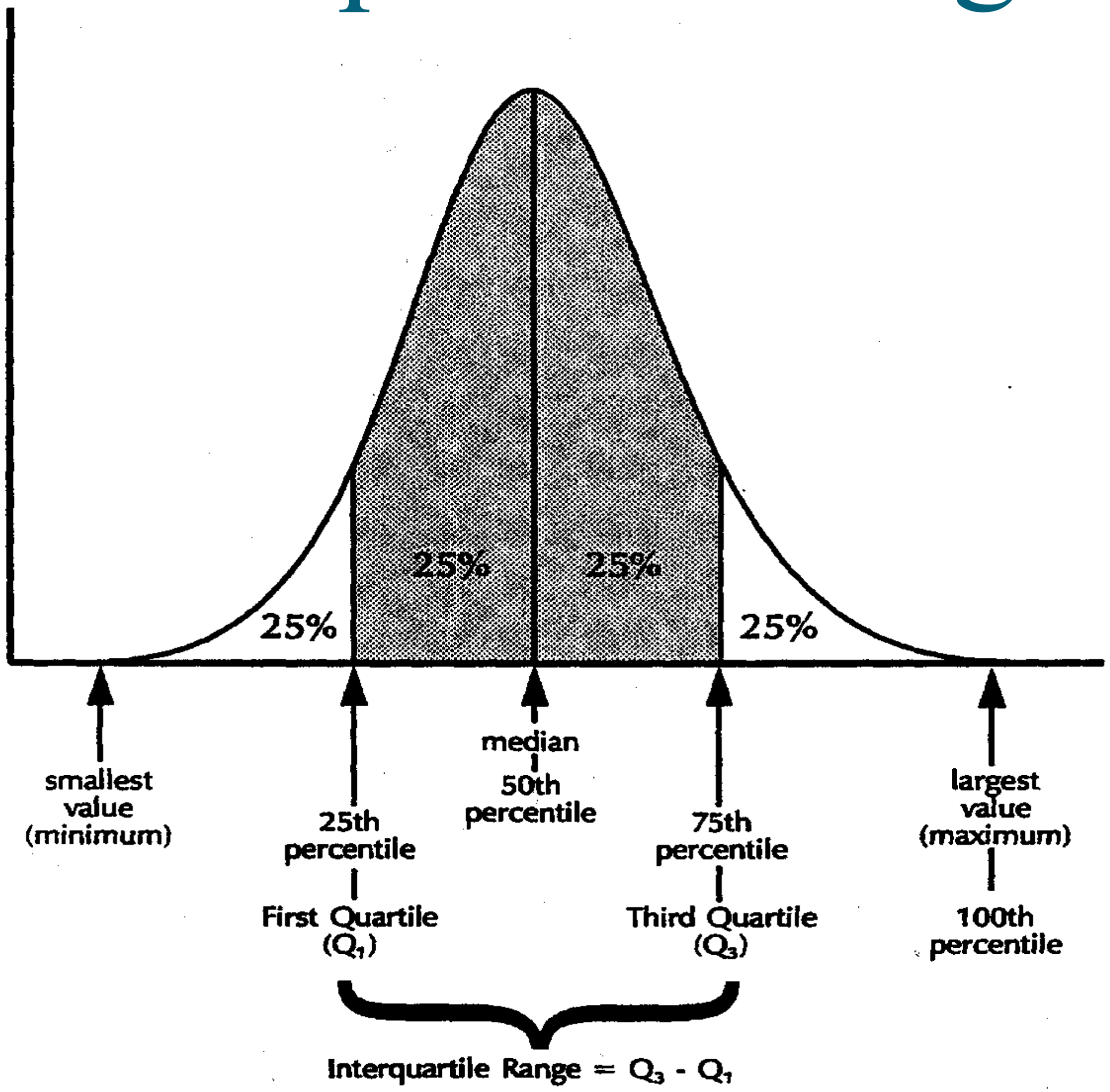
Inter quartile :

IQR = $Q_3 - Q_1$

Inter quartile Range

- The inter quartile range is Q_3-Q_1
- 50% of the observations in the distribution are in the inter quartile range.
- The following figure shows the interaction between the quartiles, the median and the inter quartile range.

Inter quartile Range



Percentiles and Quartiles

- Maximum is 100th percentile: 100% of values lie at or below the maximum

- Median is 50th percentile: 50% of values lie at or below the median

- Any percentile can be calculated. But the most common are 25th (1st Quartile) and 75th (3rd Quartile)

Table 3.3 Cumulative percentages for different ranges of haemoglobin levels of 70 women.

Observation	Cumulative percentage	Haemoglobin level (g/100 ml)		Quartile
1	1.4	8.8	Minimum = 8.8	1
2	2.9	9.3		1
3	4.3	9.4		1
4	5.7	9.7		1
5	7.1	10.2		
⋮	⋮	⋮		
15	21.4	10.8		1
16	22.9	10.9		1
17	24.3	10.9	Lower quartile = 10.9	1
18	25.7	10.9		1
19	27.1	11.0		2
20	28.6	11.0		2
⋮	⋮	⋮		
33	47.1	11.7		2
34	48.6	11.8		2
35	50.0	11.8	Median = 11.85	2
36	51.4	11.9		3
37	52.9	11.9		3
38	54.3	12.0		3
⋮	⋮	⋮		
50	71.4	12.9		3
51	72.9	12.9		3
52	74.3	13.0		3
53	75.7	13.1	Upper quartile = 13.1	4
54	77.1	13.1		4
55	78.6	13.2		4
⋮	⋮	⋮		
66	94.3	14.6		4
67	95.7	14.6		4
68	97.1	14.7		4
69	98.6	14.9		4
70	100	15.1	Maximum = 15.1	4

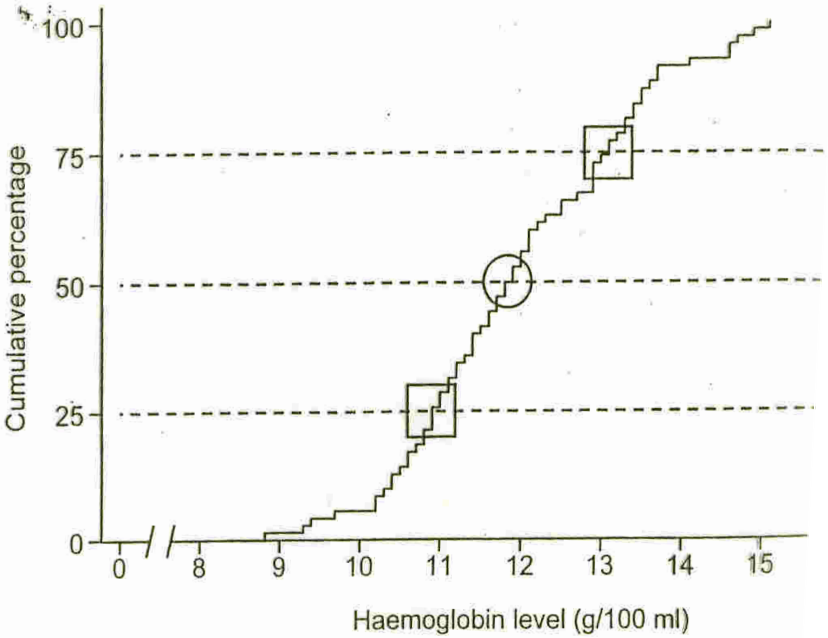
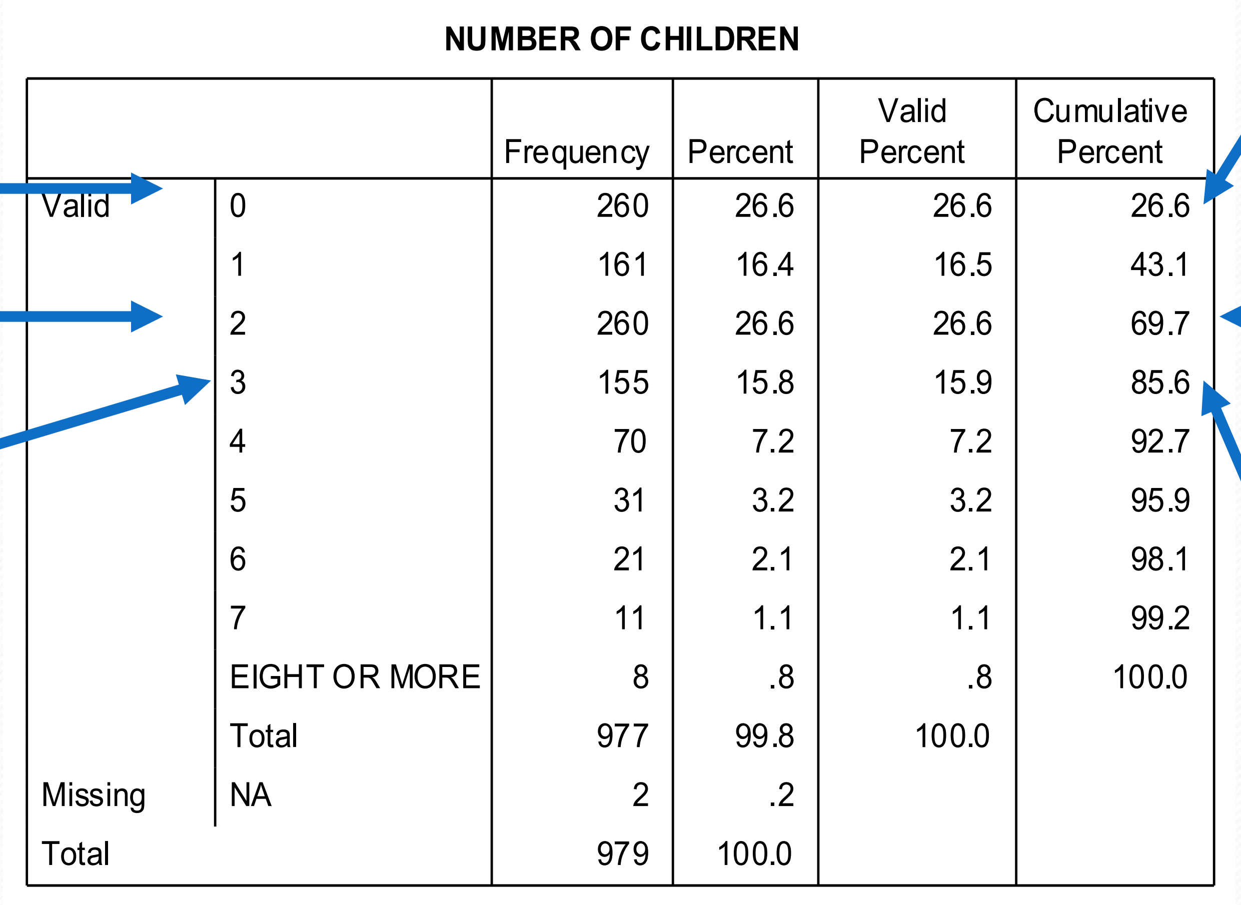


Fig. 3.7 Cumulative frequency distribution of haemoglobin levels of 70 women, with the median m
a circle, and lower and upper quartiles marked by squares.

Locating Percentiles in a Frequency Distribution

- A percentile is a score below which a specific percentage of the distribution falls(the median is the 50th percentile.
- The 75th percentile is a score below which 75% of the cases fall.
- The median is the 50th percentile: 50% of the cases fall below it
- Another type of percentile :The quartile lower quartile is 25th percentile and the upper quartile is the 75th percentile



NUMBER OF CHILDREN

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	260	26.6	26.6	26.6
	1	161	16.4	16.5	43.1
	2	260	26.6	26.6	69.7
	3	155	15.8	15.9	85.6
	4	70	7.2	7.2	92.7
	5	31	3.2	3.2	95.9
	6	21	2.1	2.1	98.1
	7	11	1.1	1.1	99.2
	EIGHT OR MORE	8	.8	.8	100.0
	Total	977	99.8	100.0	
Missing	NA	2	.2		
Total		979	100.0		

25th percentile

50th percentile

80th percentile

25% included here

50% included here

80% included

here

VARIANCE

Deviations of each observation from the mean, then averaging the sum of squares of these deviations.

STANDARD DEVIATION

“ ROOT- MEANS-SQUARE-DEVIATIONS”

Standard Deviation

- To “undo” the squaring of difference scores, take the square root of the variance.

- Return to original units rather than squared units.

Quantifying Uncertainty

Standard deviation: measures the variation of a variable in the sample.

-Technically,

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

Example

Data: X = {6, 10, 5, 4, 9, 8}; N = 6

X	$X - \bar{X}$	$(X - \bar{X})^2$
6	-1	1
10	3	9
5	-2	4
4	-3	9
9	2	4
8	1	1
Total: 42		Total: 28

Mean:

$$\bar{X} = \frac{\sum X}{N} = \frac{42}{6} = 7$$

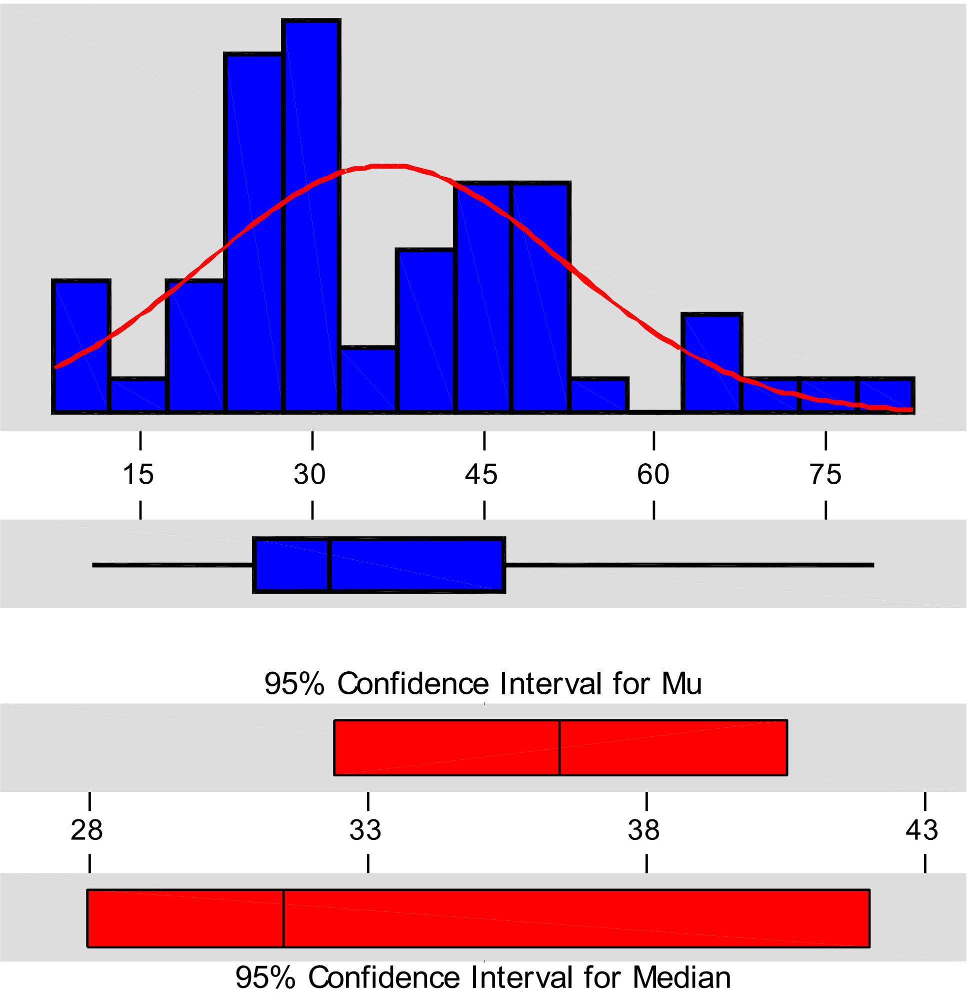
Variance:

$$s^2 = \frac{\sum (\bar{X} - X)^2}{N} = \frac{28}{6} = 4.67$$

Standard Deviation:

$$s = \sqrt{s^2} = \sqrt{4.67} = 2.16$$

Descriptive Statistics



Variable: Age

Anderson-Darling Normality Test

A-Squared:	0.962
P-Value:	0.014
Mean	36.4500
StDev	15.7356
Variance	247.608
Skewness	0.679626
Kurtosis	8.51E-02
N	60
Minimum	11.0000
1st Quartile	25.0000
Median	31.5000
3rd Quartile	46.7500
Maximum	79.0000

95% Confidence Interval for Mu

32.3851 40.5149

95% Confidence Interval for Sigma

13.3380 19.1921

95% Confidence Interval for Median

28.0000 42.0000

WHICH MEASURE TO USE ?

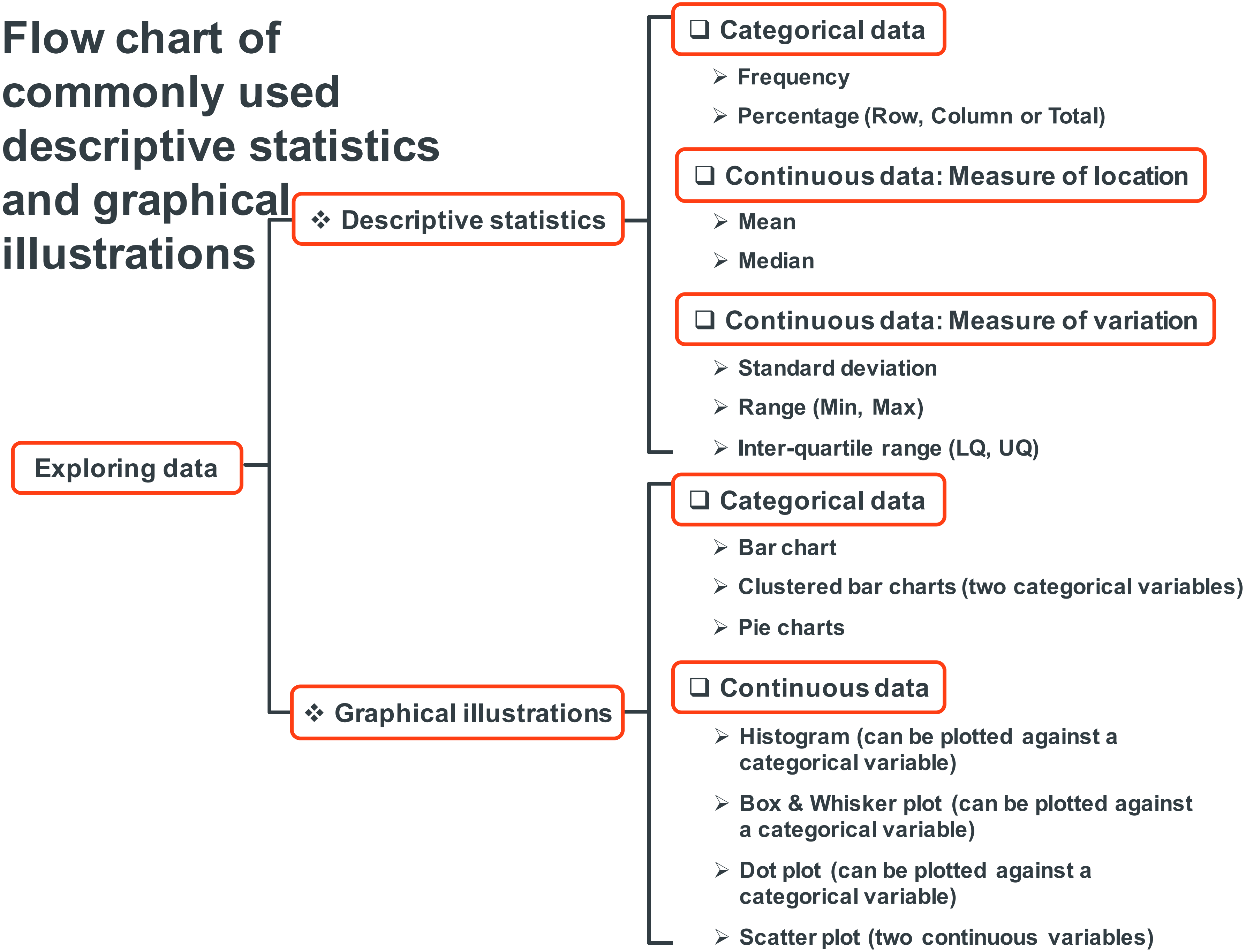
DISTRIBUTION OF DATA IS SYMMETRIC

---- USE MEAN & S.D.,

DISTRIBUTION OF DATA IS SKEWED

---- USE MEDIAN & QUARTILES

Flow chart of commonly used descriptive statistics and graphical illustrations



Exploring data

- ❖ Descriptive statistics
  - ❑ Categorical data
    - ➢ Frequency
    - ➢ Percentage (Row, Column or Total)
  - ❑ Continuous data: Measure of location
    - ➢ Mean
    - ➢ Median
  - ❑ Continuous data: Measure of variation
    - ➢ Standard deviation
    - ➢ Range (Min, Max)
    - ➢ Inter-quartile range (LQ, UQ)
- ❖ Graphical illustrations
  - ❑ Categorical data
    - ➢ Bar chart
    - ➢ Clustered bar charts (two categorical variables)
    - ➢ Pie charts
  - ❑ Continuous data
    - ➢ Histogram (can be plotted against a categorical variable)
    - ➢ Box & Whisker plot (can be plotted against a categorical variable)
    - ➢ Dot plot (can be plotted against a categorical variable)
    - ➢ Scatter plot (two continuous variables)