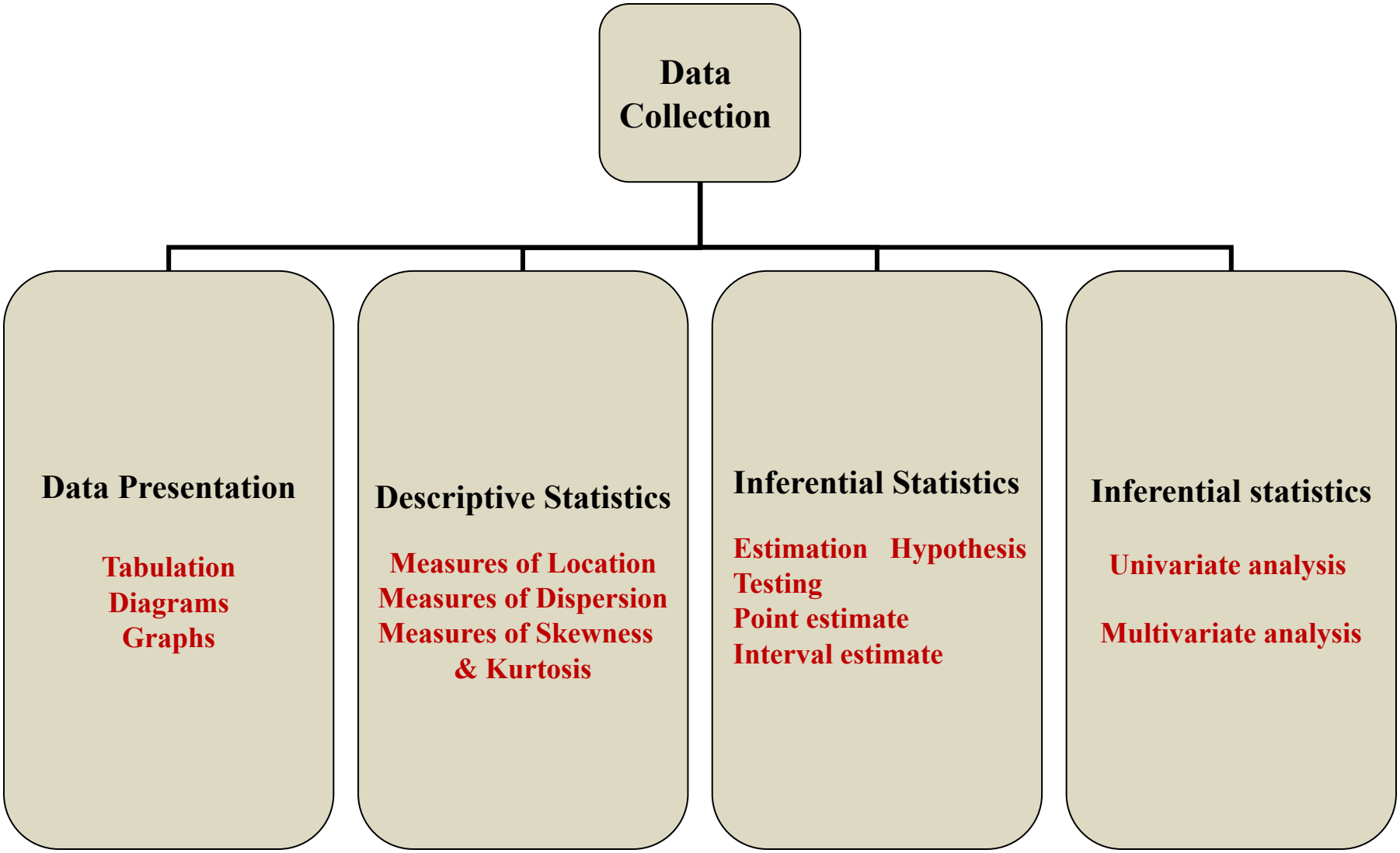
Description of Data (Using Summary & Variability measures)

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

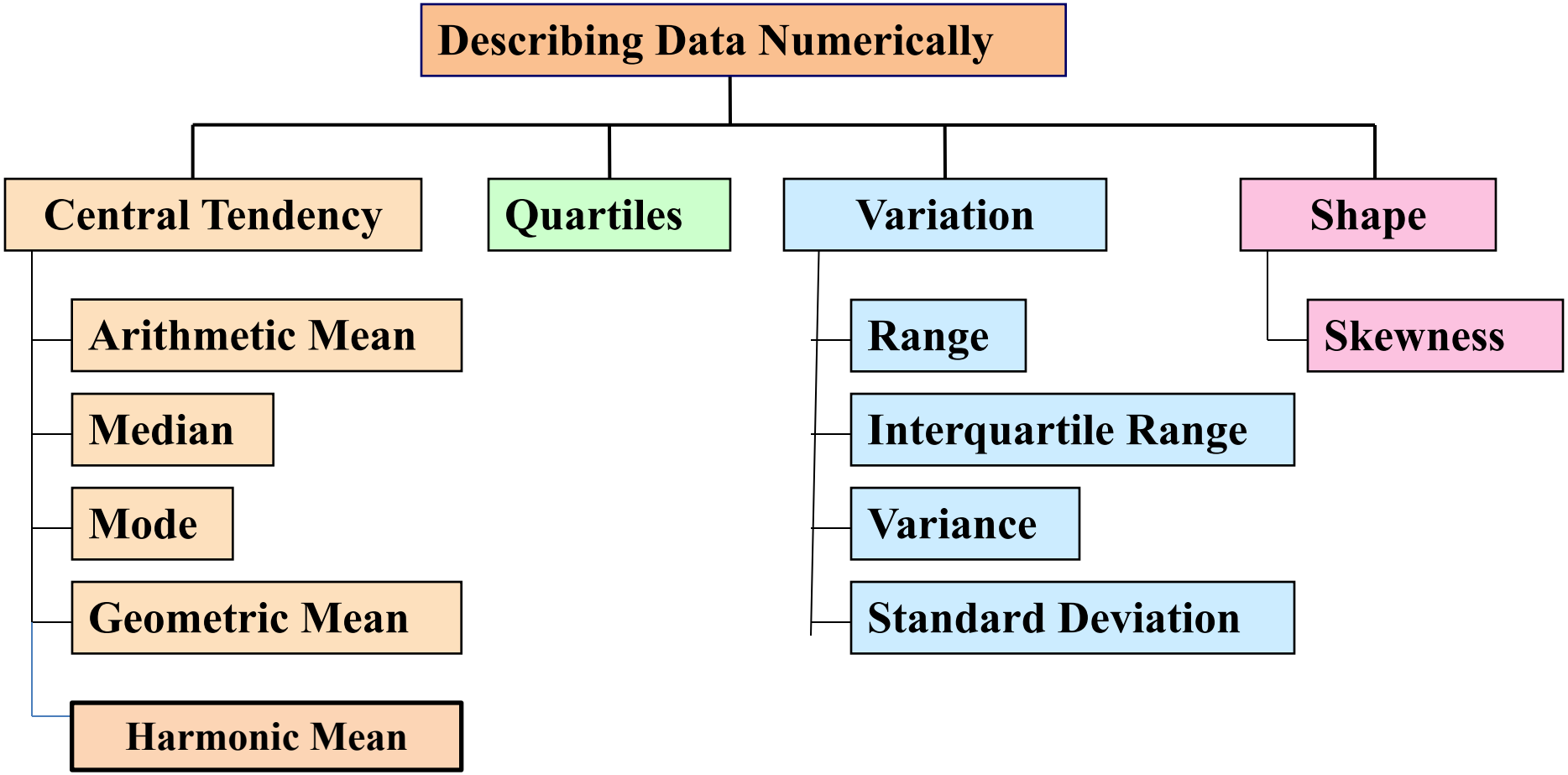
Objectives of this session

- 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.

Investigation



Summary & Variability Measures



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

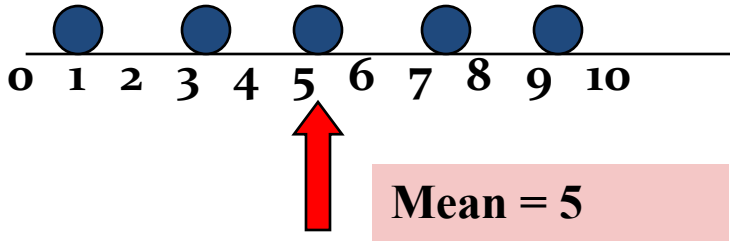
- Sum the scores (ΣX):
  $1 + 5 + 4 + 3 + 2 = 15$

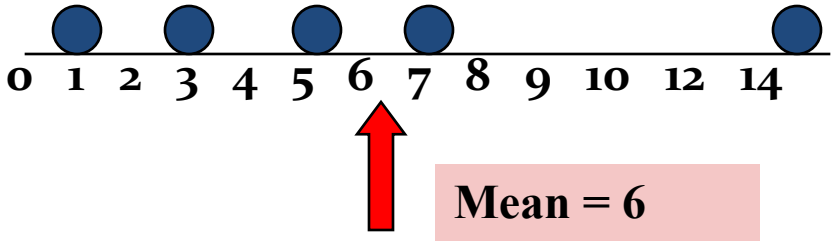
- Divide the sum ($\Sigma X = 15$) by the number of scores ($N = 5$): $15 / 5 = 3$

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

Mean (Arithmetic Mean) *(continued)*

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





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

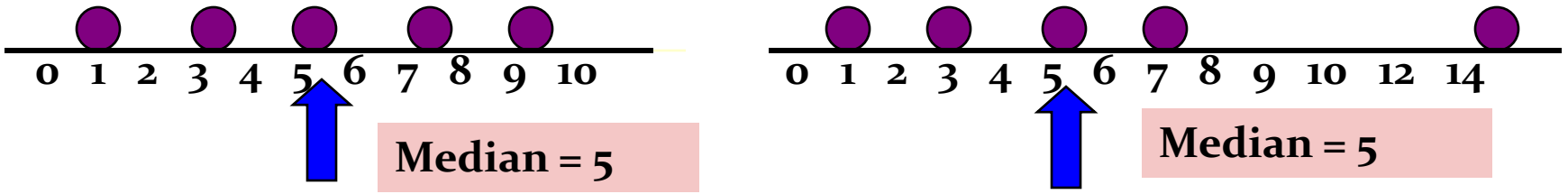
- Median = average of 3rd and 4th scores:
  (19 + 18) / 2 = 18.5

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



- 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)

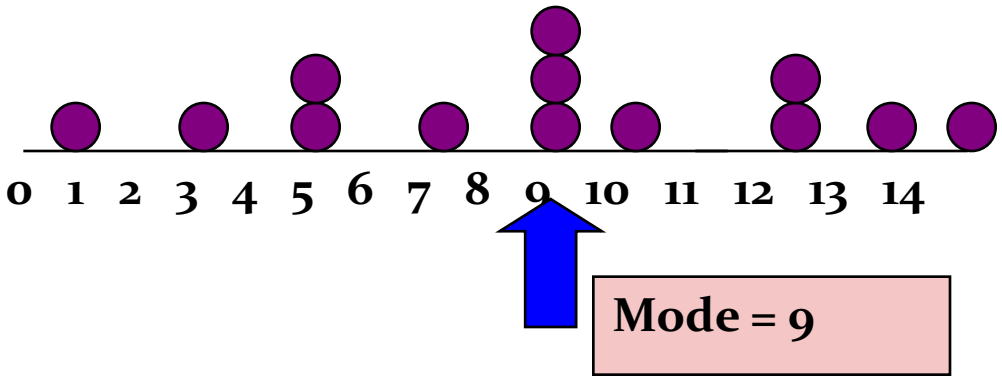
e.g. 1 2 3 4 5 6 7 8 9 100

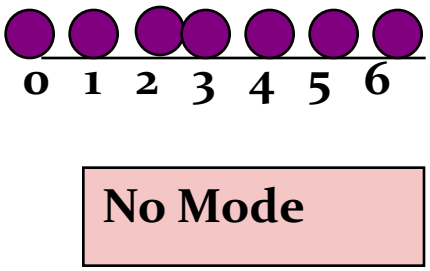
Mean = 14.5 (median = 5.5)

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





The Shape of Distributions

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

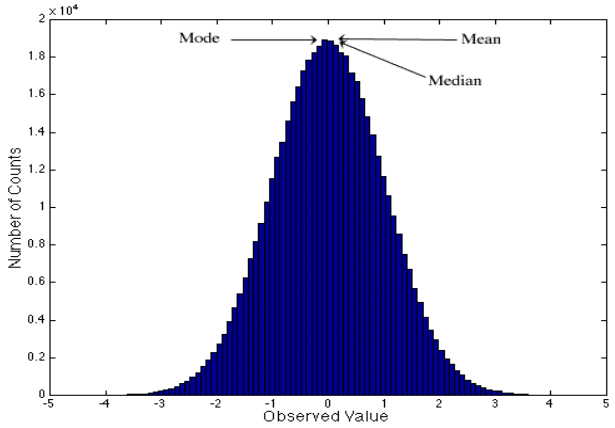
Symmetrical Distributions

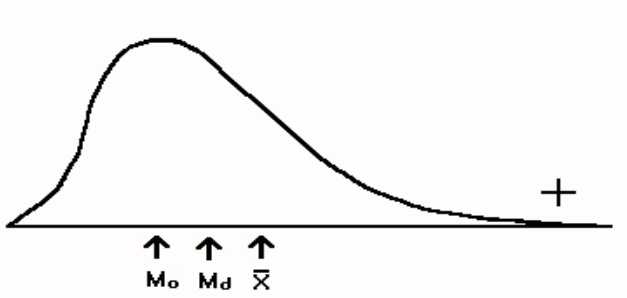
- A distribution is symmetrical if the frequencies at the right and left tails of the distribution are identical, so that if it is divided into two halves, each will be the mirror image of the other.

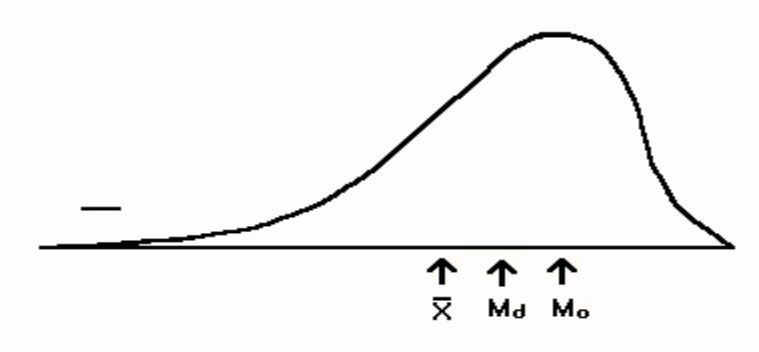
- **In a symmetrical distribution the mean, median, and mode are identical**.

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)

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.

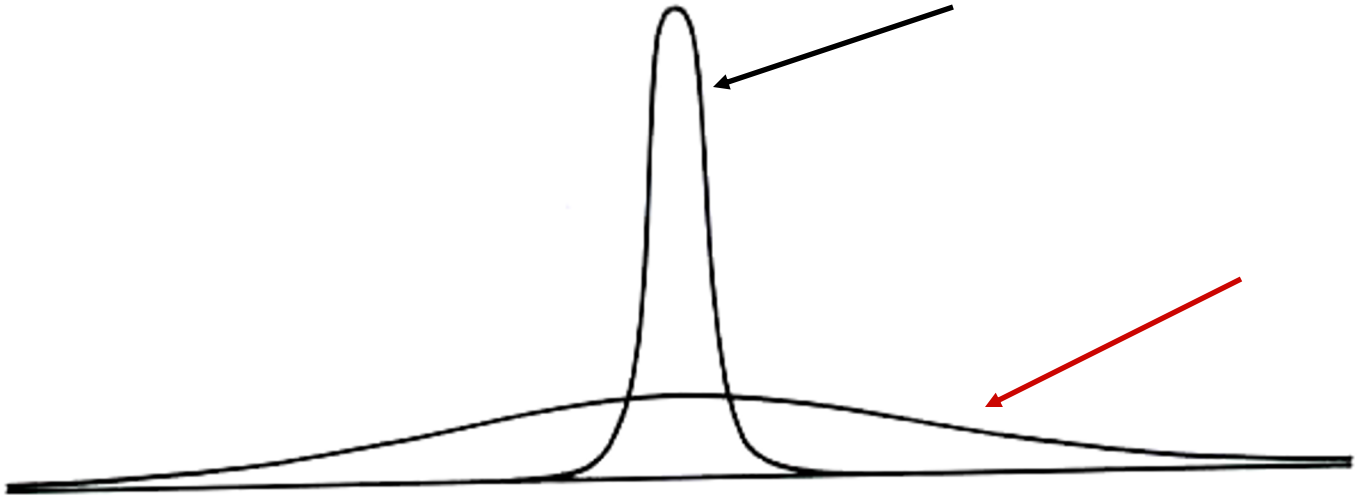
Series I: 70 70 70 70 70 70 70 70 70 70

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

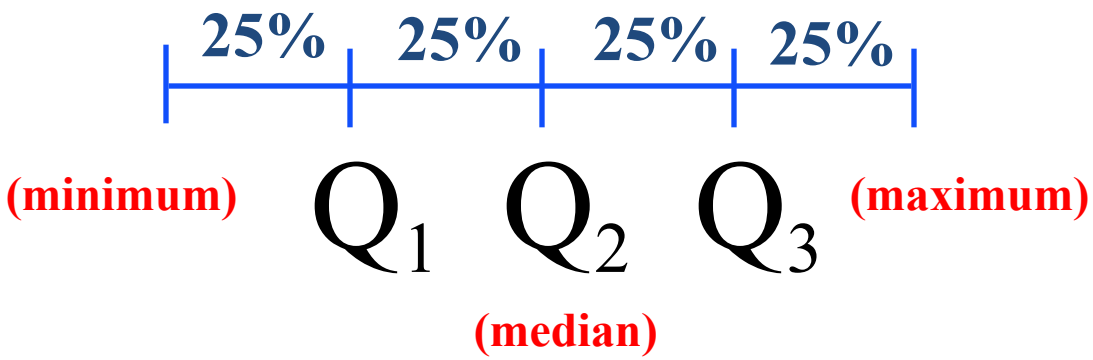
  52, 76, 100, 36, 86, 96, 20, 15, 57, 64, 64, 80, 82, 83, 30, 31, 31, 31, 32, 37, 38, 38, 40, 40, 41, 42, 47, 48, 63, 63, 72, 79, 70, 71, 89

- Range: 100-15 = 85

Quartiles

Q_1, Q_2, Q_3

divides ranked scores into four equal parts



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

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

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

Inter quartile :

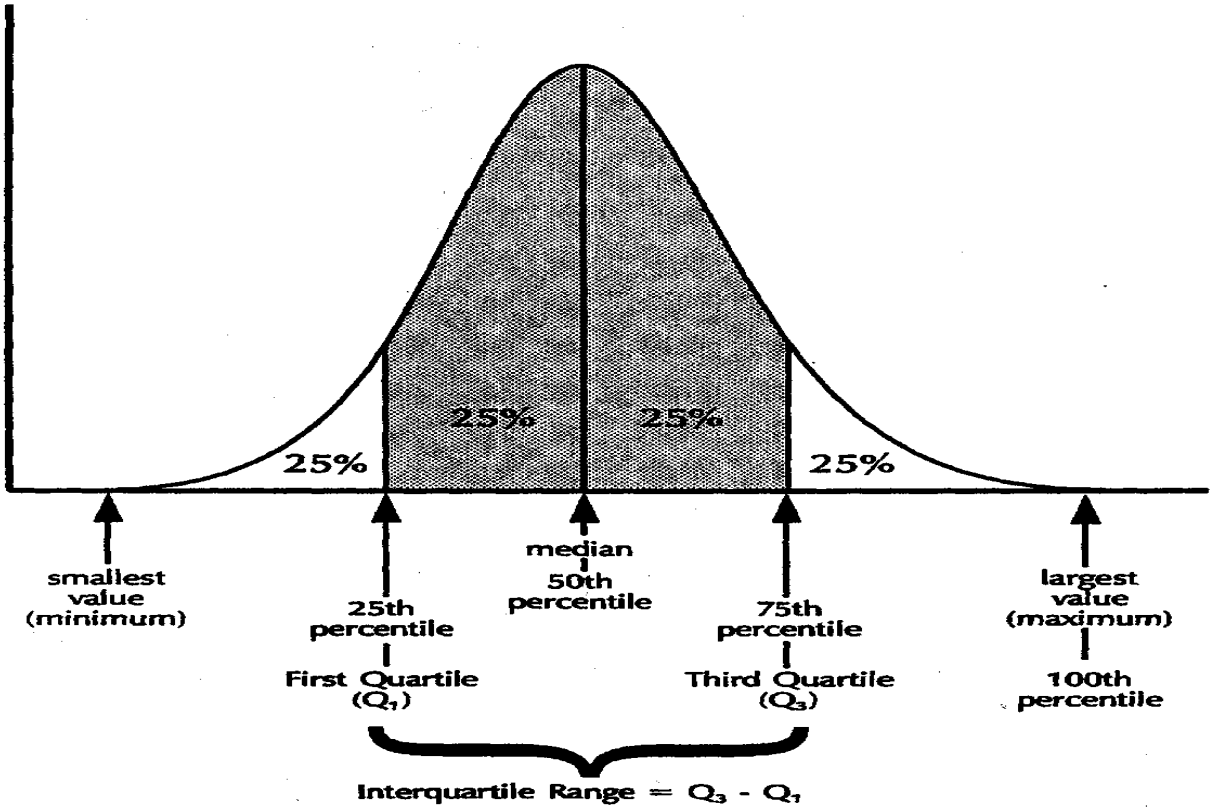
$$IQR = Q_3 - Q_1$$

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

FIGURE 3.8
The middle half of the observations in a frequency distribution lie within the interquartile 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 25^{th} (1^{st} Quartile) and 75^{th} (3^{rd} Quartile)

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,

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

$X \qquad X - \overline{X} \qquad (X - \overline{X})^2$

Calculation of Variance & Standard deviation

- Using the deviation & computational method to calculate the variance and standard deviation
- Example: 3,4,4,4,6,7,7,8,8,9 ; Given n=10; Sum= 60; Mean = 6

$$S = \sqrt{\frac{\sum (X - \overline{X})^2}{n}}$$

$$S = \sqrt{\frac{(3-6)^2 + (4-6)^2 + (4-6)^2 + (4-6)^2 + (6-6)^2 + (7-6)^2 + (7-6)^2 + (8-6)^2 + (8-6)^2 + (9-6)^2}{10}}$$

$$S = \sqrt{\frac{40}{10}} = 2.0; \text{var}\,iance = 4$$

X	X²
3	9
4	16
4	16
4	16
6	36
7	49
7	49
8	64
8	64
9	81
Sum: 60	**Sum: 400**

$$S = \sqrt{\frac{n\sum X^2 - (\sum X)^2}{n^2}}$$

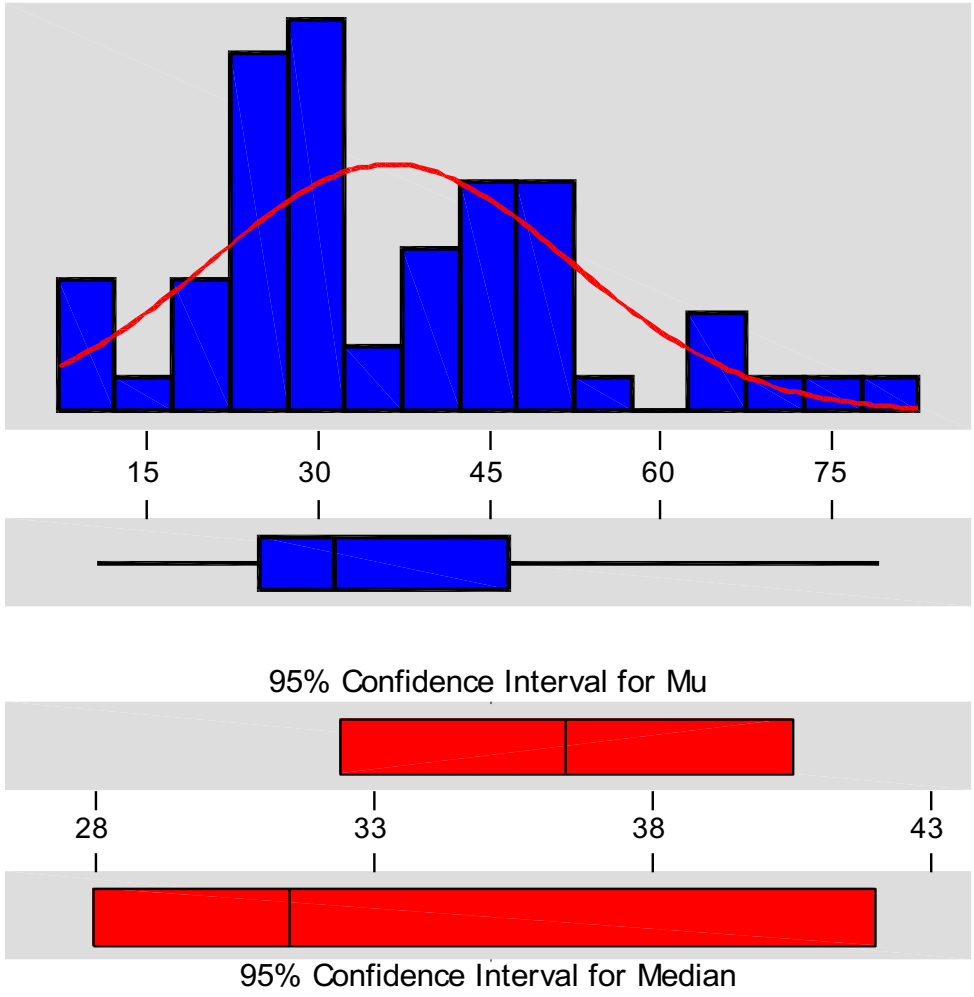
$$S = \sqrt{\frac{10(400) - (60)^2}{10^2}}$$

$$S = \sqrt{\frac{4000 - 3600}{100}}$$

$$S = \sqrt{4.0}$$

$$S = 2.0, \text{var}\,iance = 4$$

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