15

DESCRIPTION OF DATA

KSU COLLEGE OF MEDICINE
2019 - 2020

ACKNOWLEDGMENTS

TEAM MEMBERS

ROTANA KHATEEB

ABDULLAH ALESSA

REVIEWER

YAZEED AL-DOSSARE

TABLE OF **CONTENTS**

Investigation

Median Examples

WHICH MEASURE TO USE ?

Skewed

Quartiles

LECTURE **OBJECTIVES**

By the end of this lecture, I am able to:

- 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

Data Presentation

Tabulation
Diagrams
Graphs

Descriptive Statistics

Measures of Location
Measures of Dispersion
Measures of Skewness & Kurtosis

Inferential Statistics

Estimation
Testing
Point estimate
Interval estimate
Hypothesis

Inferential statistics

Univariate analysis

Multivariate analysis

Summary & Variability Measures

Describing Data Numerically

Central Tendency

Arithmetic Mean
Median
Mode
Geometric Mean
(rarely used)
Harmonic Mean
(rarely used)

Variation

Deviation
Range
Standard Deviation
Variance

Quartiles

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.

Three common measures of central tendency:

The Mean	The Median	The Mode
sum/total (average)	middle number of all data	most frequent value

The Mean

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

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

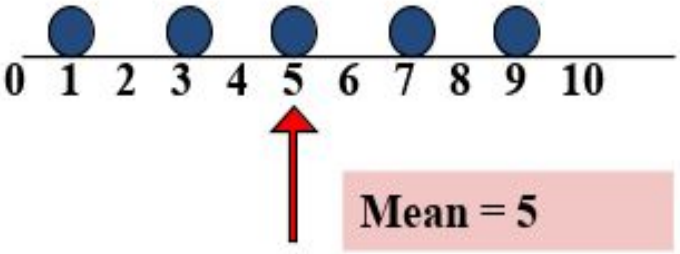
Divide the sum (ΣX = 15) by the number of scores (N = 5): 15 / 5 = 3

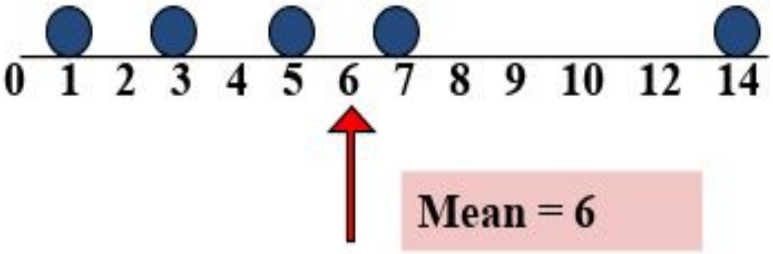
Mean affected by extreme value = X = 3

Mean (Arithmetic Mean)

The most common measure of central tendency

Affected by extreme values (outliers)*
extreme value mean for example is the student mark is between 10-12 but there is 4 student get full mark or the opposite if there is 3 student get zero in the exam





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

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 Examples

Example 1:

- 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
-first we arrange the number from high to low
-N= sample size
-here the median because the sample size is double (even) we take the average of the middle two numbers
'the median not affected by extreme values'

Example 2:

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

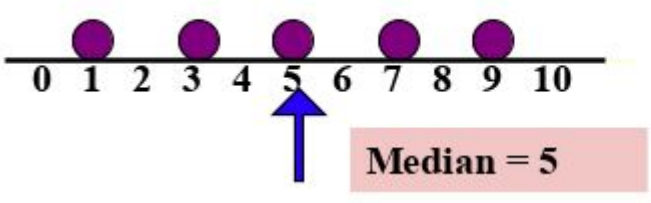
Sort the scores:
15 14 12 10 10 9 8 8 7 3 3
Determine the middle score:
middle = (N + 1) / 2 = (11 + 1) / 2 = 6
Middle score = median = 9
the only difference here is that the sample size is single (odd) so after we rearrange them we take the middle one

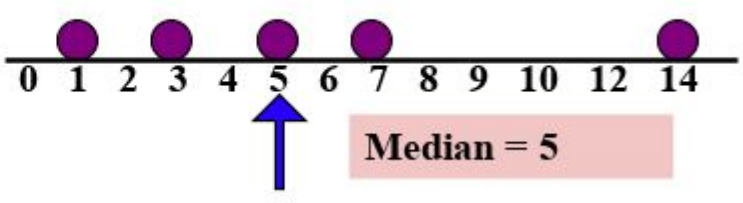
Median

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 (example if n=42 then the median is the average of the 21st and 22nd values)





Not affected by extreme values

Measures of Central Tendency

- 1 2 3 4 5 6 7 8 9 10

Mean	Median
5.5	5.5

- 1 2 3 4 5 6 7 8 9 20

6.5	5.5

- 1 2 3 4 5 6 7 8 9

5.5	5.5

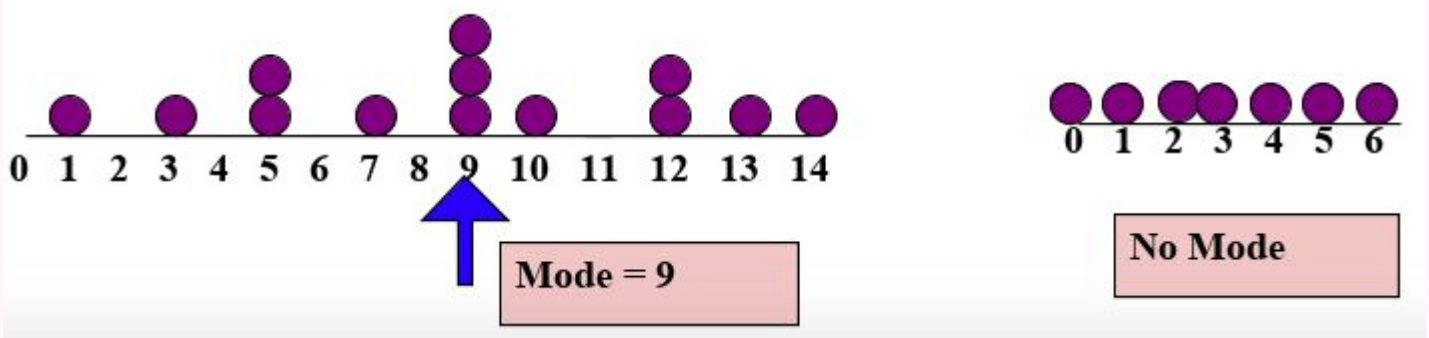
Mean is the most frequently used but is sensitive to extreme scores

here we notice that the mean is changing but the median is constant why? because the mean get affected by extreme value in eg 2 the extreme value is 20 and in eg 3 is 100

Mode

mainly applied outside educational purpose

1. *A Value that occurs most often*
2. Not affected by extreme values
3. Used for either numerical or categorical(nominal)data
4. There may be no mode (sample size isis low for eg when you take the mark for 3 student only there will be no repetid)
5. There may be several modes



The Shape of Distributions

Symmetrical

Skewed

shift to one side

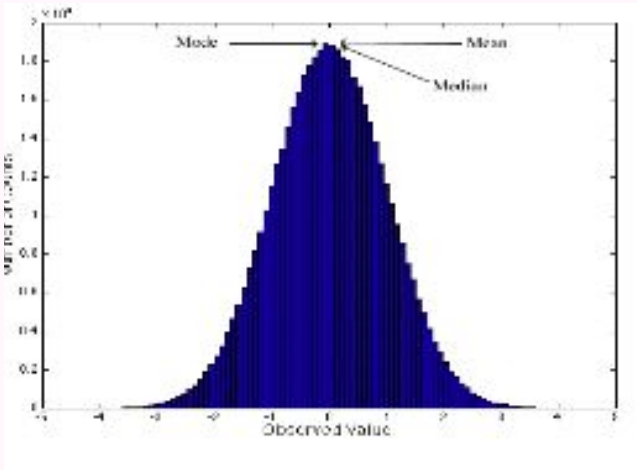
Depending on whether there are more frequencies at one end of the distribution than the other.

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.

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



Skewed:

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

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

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

Skewed

*as you go to the right the number increase

- positively (skewed to the right)

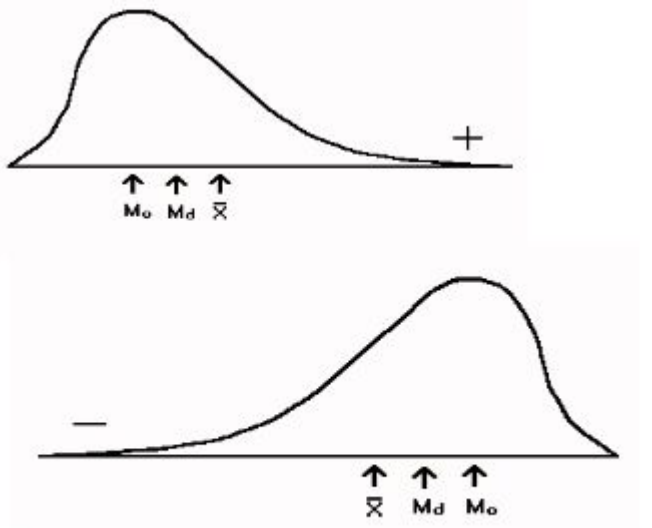
it tails off toward larger values

pic1:for example the mark of medicine only 5% of the student will get A+ but the majority will be less than that ,the PIC reflet the majority of student in the left(less number)and the student who get high mark on the right

- negatively (skewed to the left)

it tails off toward smaller values

pic2 :for example the student mark in research course the majority of the student get A A+ so they will be on the right of the chart



Choosing a Measure of Central tendency

IF variable is **Nominal** like gender	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 $n=5, \Sigma x=45, \bar{x}=45/5=9$

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

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

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

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 No variability

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

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

Measures of Variability

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

Range

Difference between the smallest and largest observations. The highest mark is 15 and the lowest is 10 the range will be 15-10=5 (not good mathematical because only rely on two values)

Interquartile Range

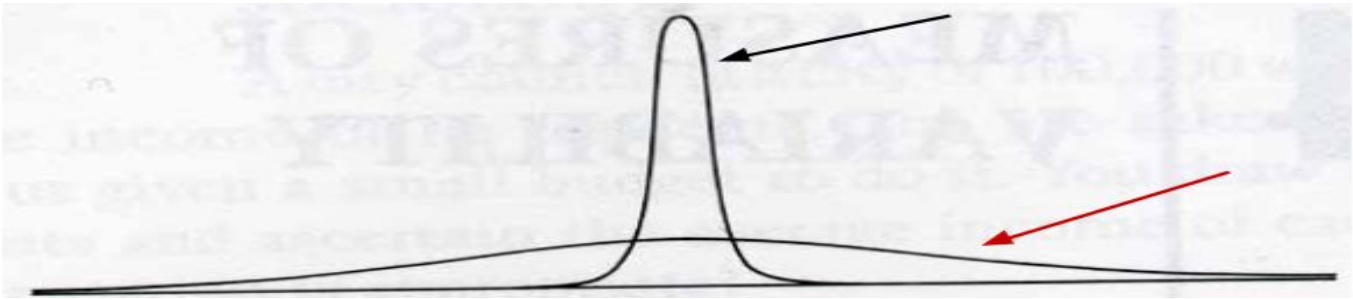
Range of the middle half of scores.

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:

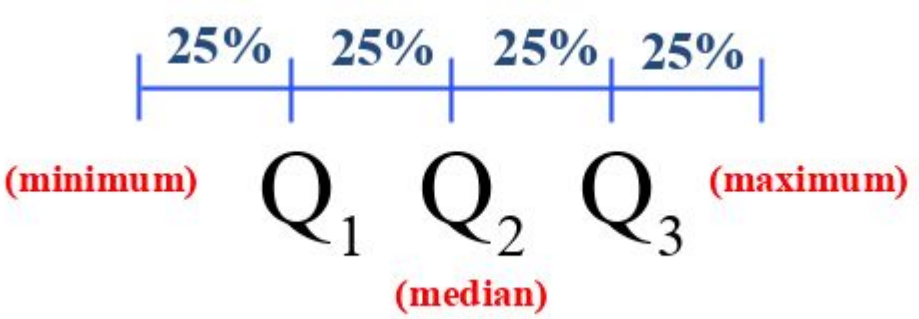
52, 76, 100, 36, 86, 96, 20, 15, 57, 64, 64, 80, 82, 83, 30, 31, 31, 31, 32, 37, 38, 38, 40, 40, 41, 42, 47, 48, 63, 63, 72, 79, 70, 71, 89

Range:

100-15 = 85

Quartiles

Q1, Q2, Q3
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 :

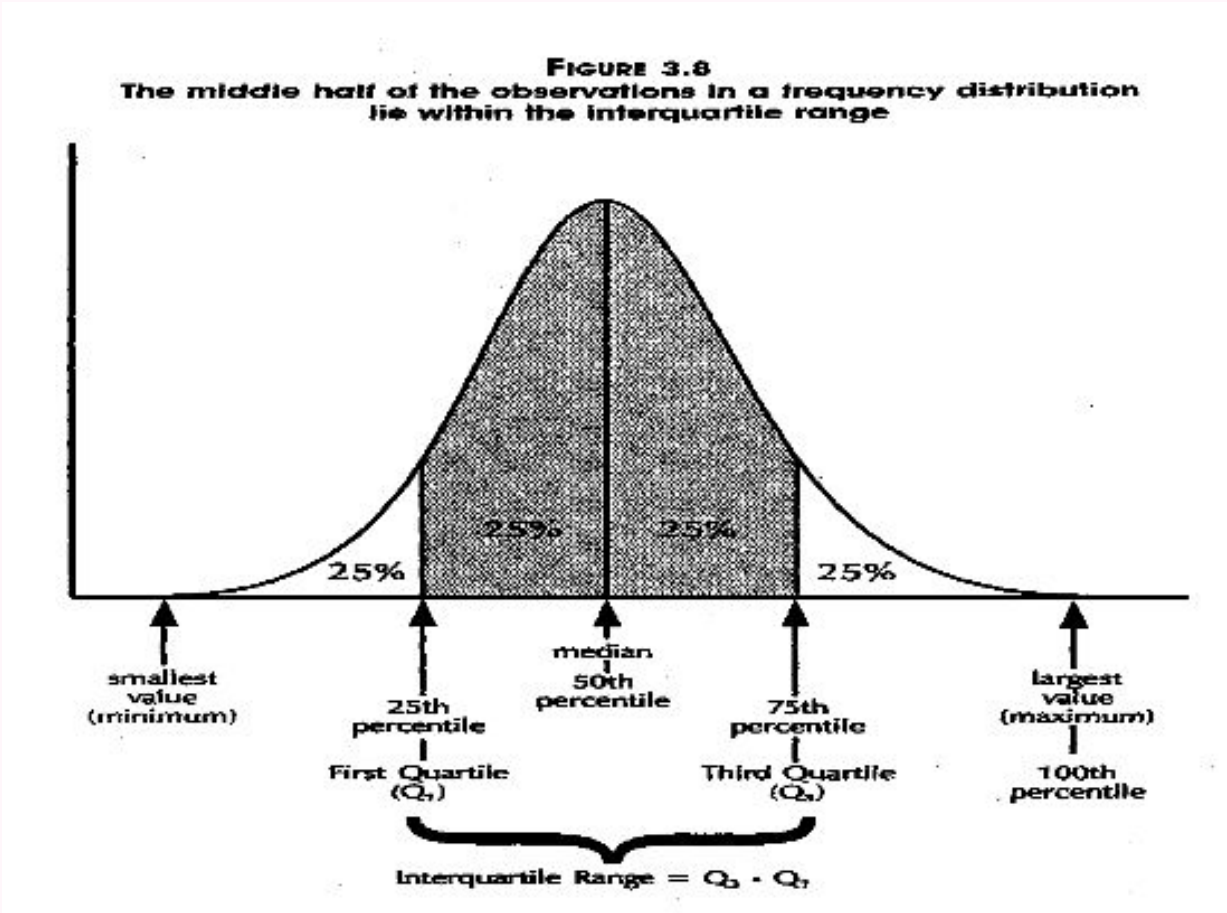
$$IQR = Q_3 - Q_1$$

Interquartile Range

The interquartile range is Q3-Q1

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 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 25th (1st Quartile) and 75th (3rd 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"

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:

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

Interpretation: All 6 values on average are deviating by 2.16. On average each student is different from other by 2.16.

Calculation of Variance & Standard Deviation

Using the deviation & computational method to calculate the variance and standard deviation

Calculation of Variance & Standard Deviation

Example

3,4,4,4,6,7,7,8,8,9

Given n=10; Sum= 60; Mean = 6

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

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

$$S = \sqrt{\frac{40}{10}} = 2.0; \text{variance} = 4$$

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**

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

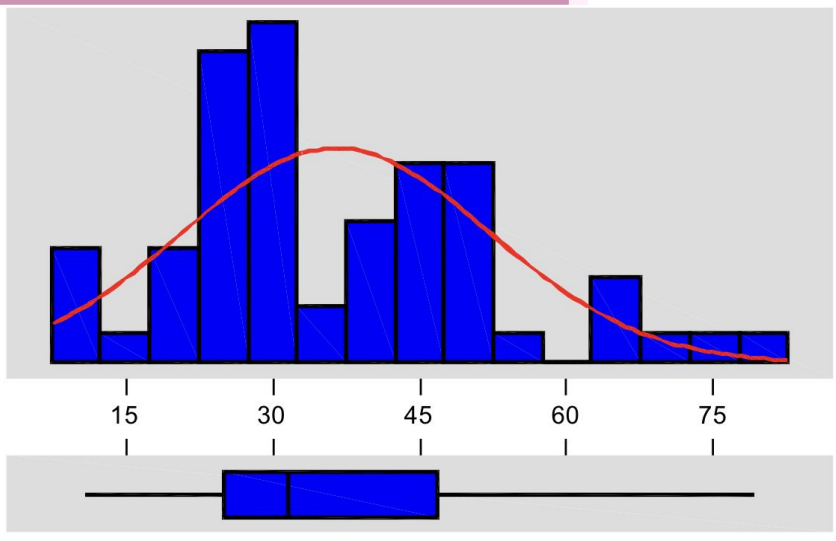
$$S = \sqrt{\frac{10(400) - (60)^2}{10^2}}$$

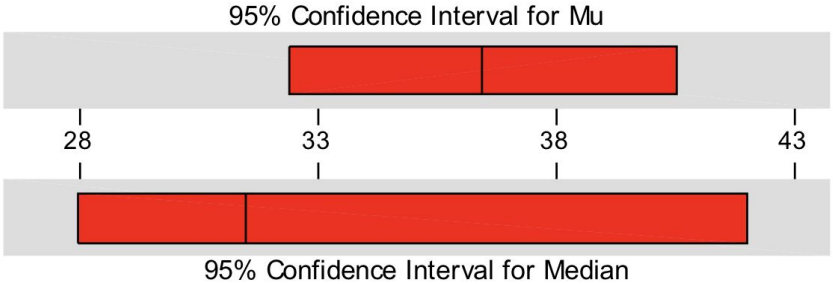
$$S = \sqrt{\frac{4000 - 3600}{100}}$$

$$S = \sqrt{4.0}$$

$$S = 2.0, \text{variance} = 4$$

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

DISTRIBUTION OF DATA IS SKEWED ----> USE MEDIAN & QUARTILES

Flow chart of commonly used descriptive statistics and graphical illustrations

- **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)
- **Exploring data**
- **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)
    - Stem and Leaf plot
    - Scatter plot (two continuous variables)