# Introduction to medical statistics

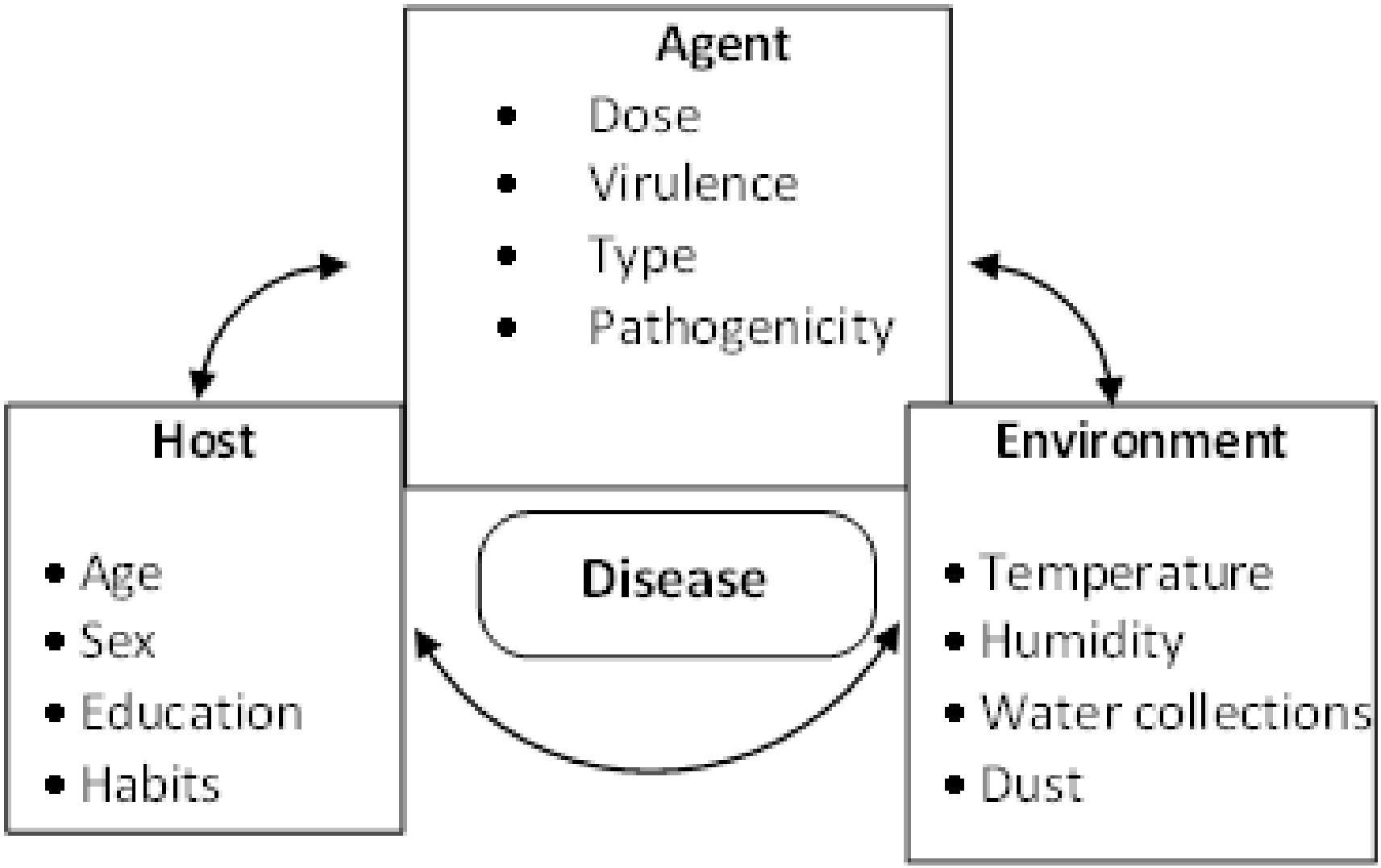
- ***The statistical science*** is a collection of rules and lows that guide the workers in any field to solve their problems in an accurate way. The main problem that faces the medical man is the disease identification or diagnosis. Also, the choice of the best method of treatment and its evaluation is another problem. With the proper use of statistical methods these two medical problems could be solved accurately in a rapid and easy way.

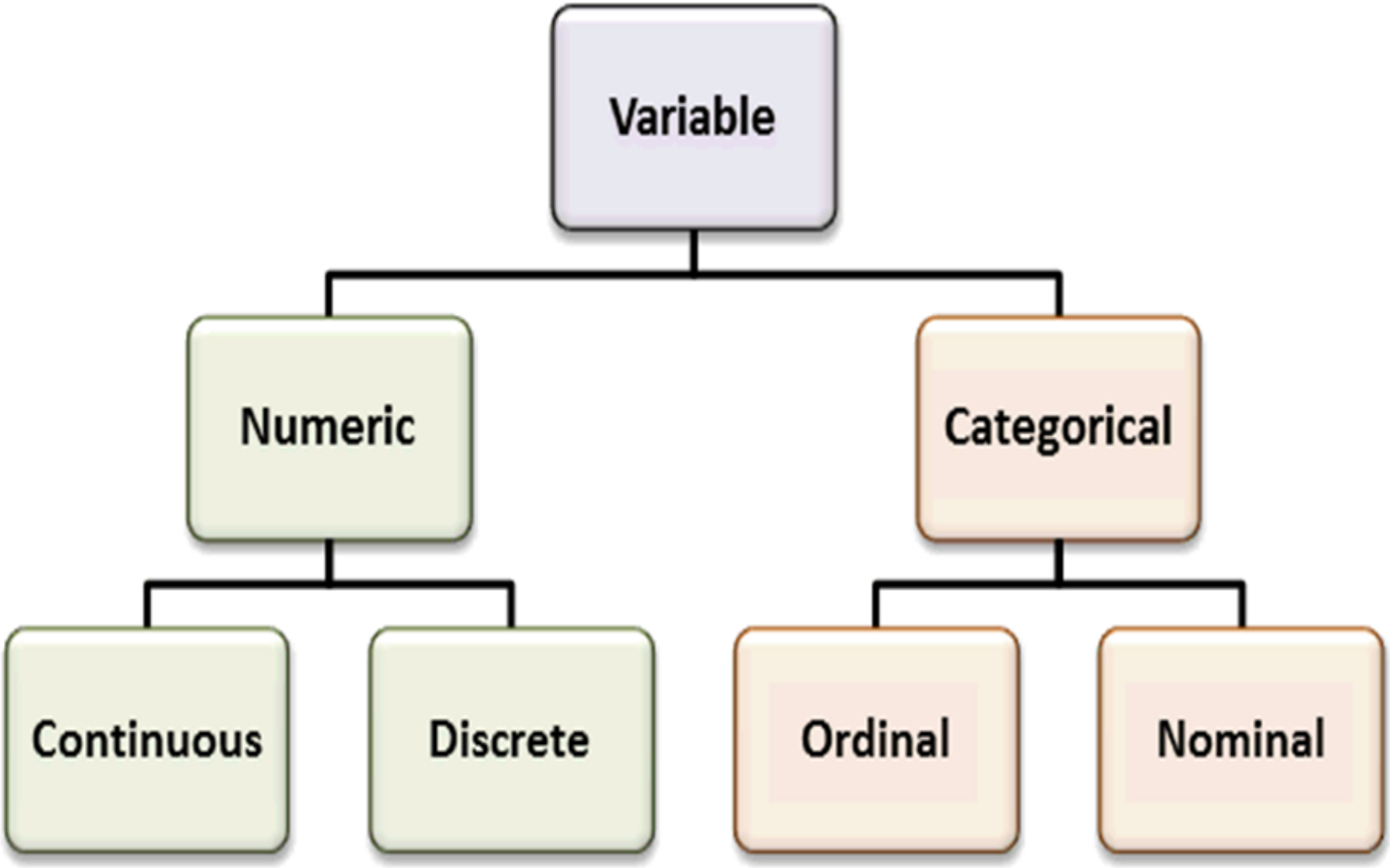
- **Statistics** as science was developed to help the collection and understanding of information in an accurate and easy way.

- ***Statistics*** as science is dealing with methods of data collection and use to suite different needs, and allow for fast, cheap and easy acquisition of data that can give the most reliable information.

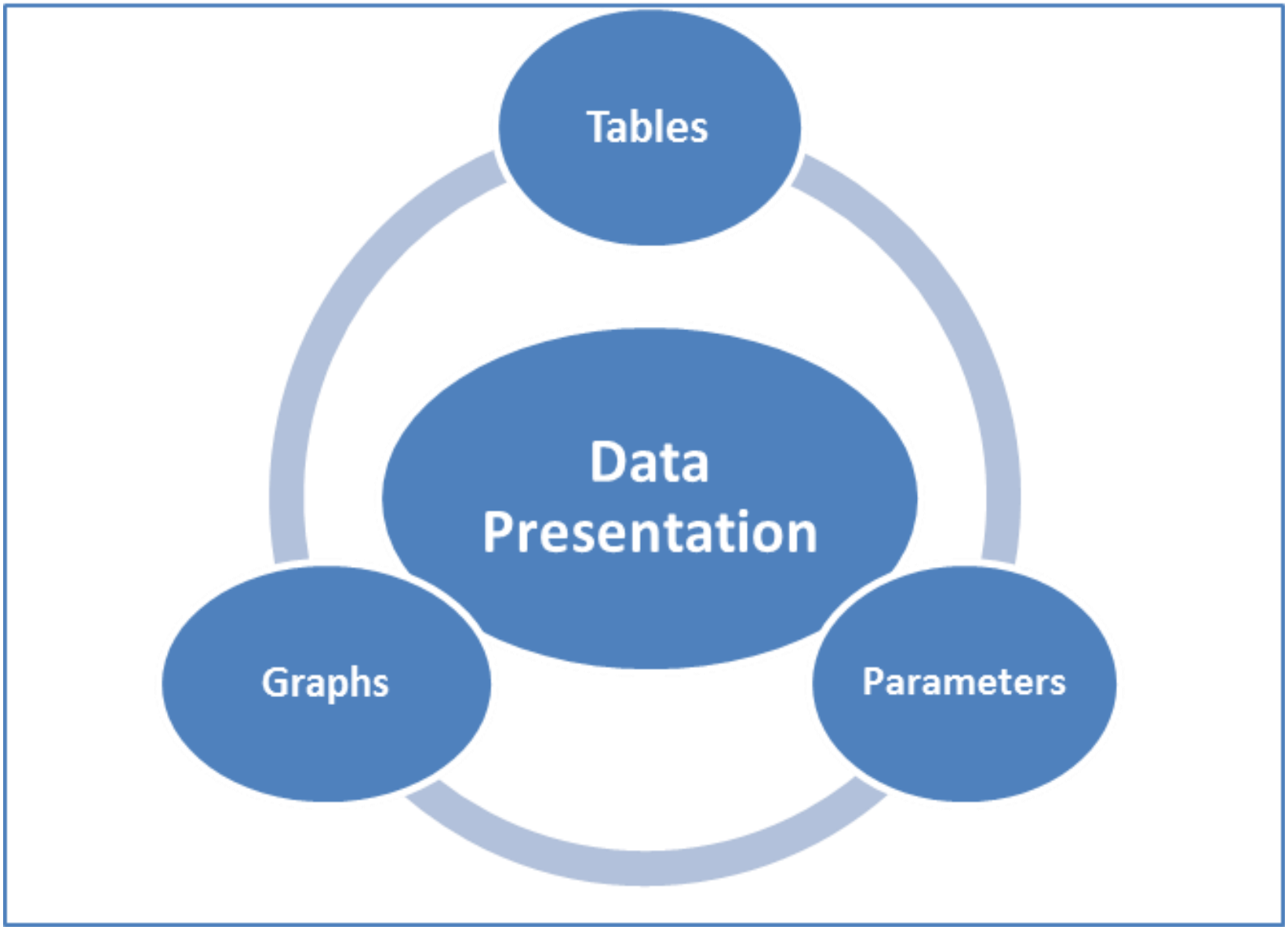
- ***Statistics*** provide rules that can suite many applications and on both small and large scale. To get the best out of the statistical use, one must apply its rules at each step of the following four main steps:

1. Data Collection.
2. Data Presentation.
3. Data Analysis.
4. Interpretation of results


Agent
Dose
Virulence
Type
Pathogenicity
Host
Age
Sex
Education
Habits
Disease
Environment
Temperature
Humidity
Water collections
Dust


Variable
Numeric
Categorical
Continuous
Discrete
Ordinal
Nominal


# Data Presentation


Tables
Data Presentation
Graphs
Parameters

- Good statistical table may have following features :

1. Table No. It is to identify the table.

2.Title of the table, whether its showing GDP, literacy etc.

3.Column and Row headings

4.Body or Main content of the table. The main data regarding the parameter.

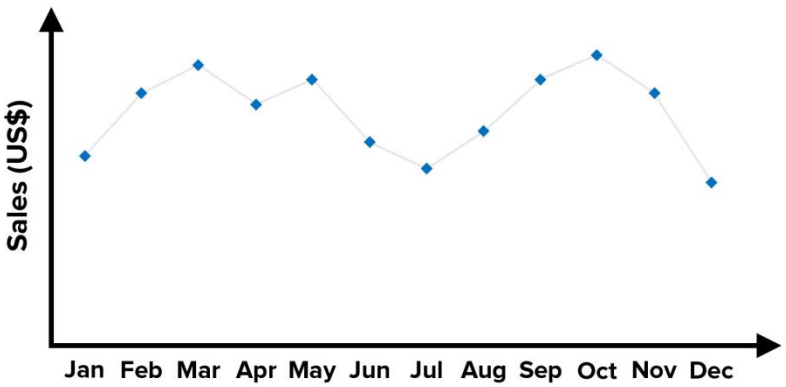
5.Unit of measure. Is the parameter measured in kms, kgs,

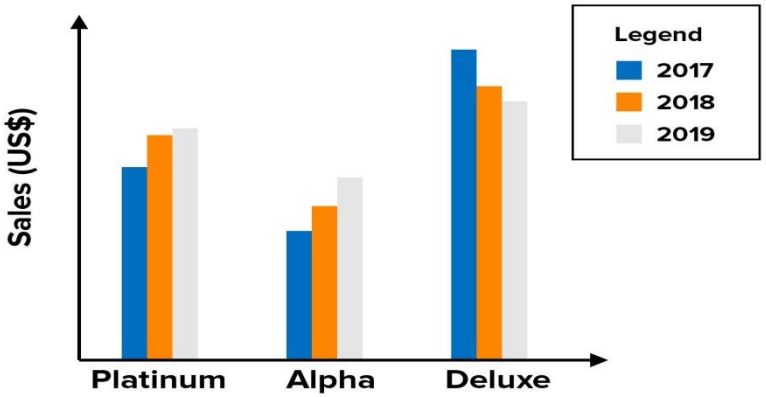
- Besides these features, a good statistical table can have a "source note" and a "footnote" describing from where the data has been taken or other information about the data.

**Sex distribution of medical statistics course students (Kasr El-Eini 2020).**

| Sex | Number | Percent |
|---|---|---|
| Males | 11 | 61.1 |
| Females | 7 | 38.9 |
| Total | 18 | 100.0 |

- Charts and graphs help to express complex data in a simple format.
- Graph  No. It is to identify the table.
- Title of the Graph , whether its showing GDP, literacy etc.




- Types are Different according to types of Data

# Parameter

- A parameter is a value that describes a characteristic of an entire population,

# Qualitative Data

1- Nominal (e.g. SEX): It is expressed as either Male or Female. i.e. just names.

2- Ordinal (e.g. Disease severity): It is expressed as Mild, Moderate or Severe. i.e. the description has a certain order in this example it is mild, moderate then severe.

# Qualitative data presentation

*Table*

**Table 2: Sex distribution of medical statistics course students (Kasr El-Eini 2020).**

| Sex | Number | Percent |
|---|---|---|
| Males | 11 | 61.1 |
| Females | 7 | 38.9 |
| Total | 18 | 100.0 |

- ***From that table you can say that males are more common than females and that they are nearly twice there number***.

- Graph A Bar B pie

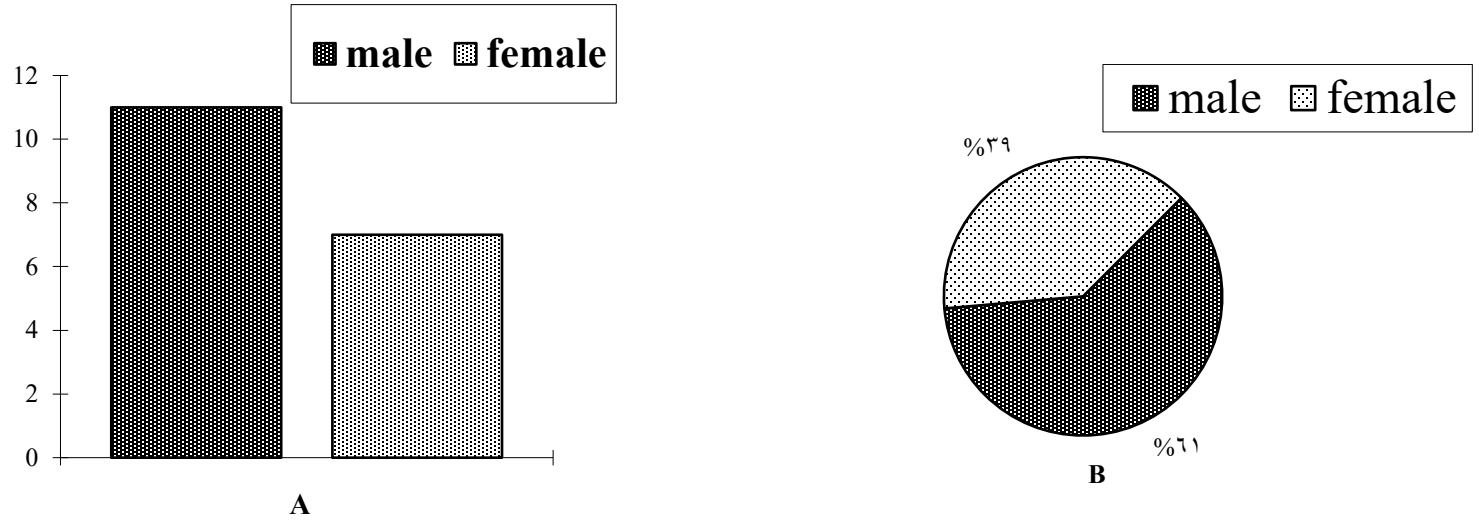



**Fig. 5: Sex distribution of medical statistics course students (Kasr El-Eini 2020).**

**Both charts show at a glance that males are more than females**

## C-Pictogram

- These are meant to convey data to the ‘man in the street’ who finds it difficult to comprehend complex charts
- Small pictures or symbols are used to present the data
- The number of pictures is proportional to the frequency size
- One picture depicts a fixed number
- A fraction of a picture can be used to represent a smaller number
- Pictograms are not effective for visualizing large sets of data In this cases, a bar chart is more effective.

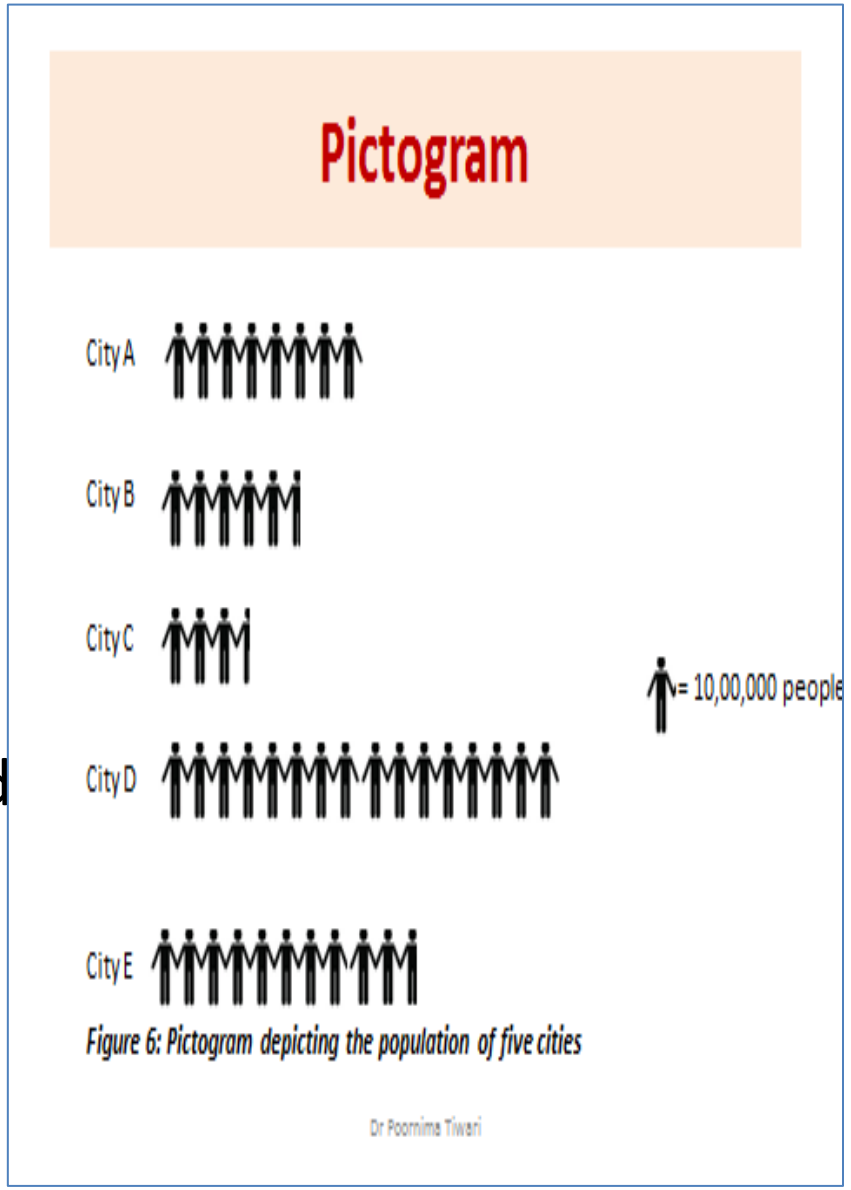
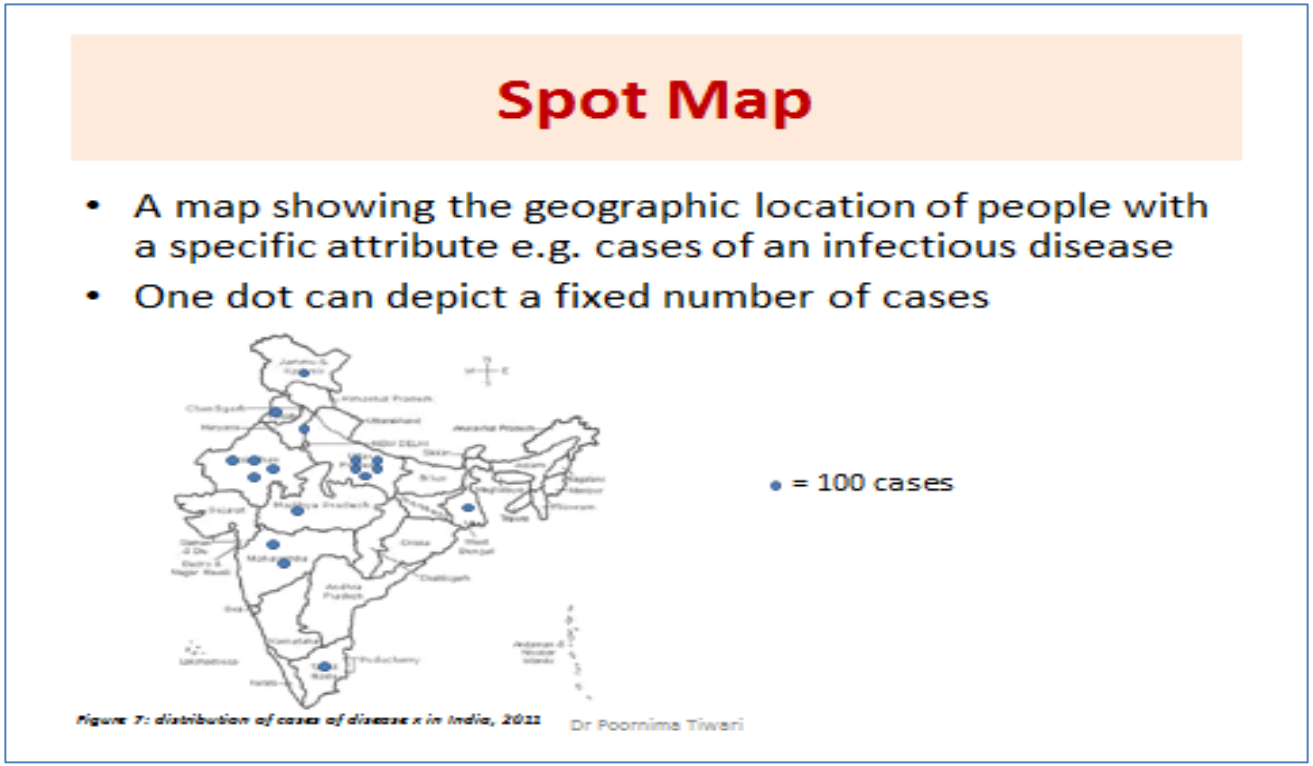


*Figure 6: Pictogram depicting the population of five cities*

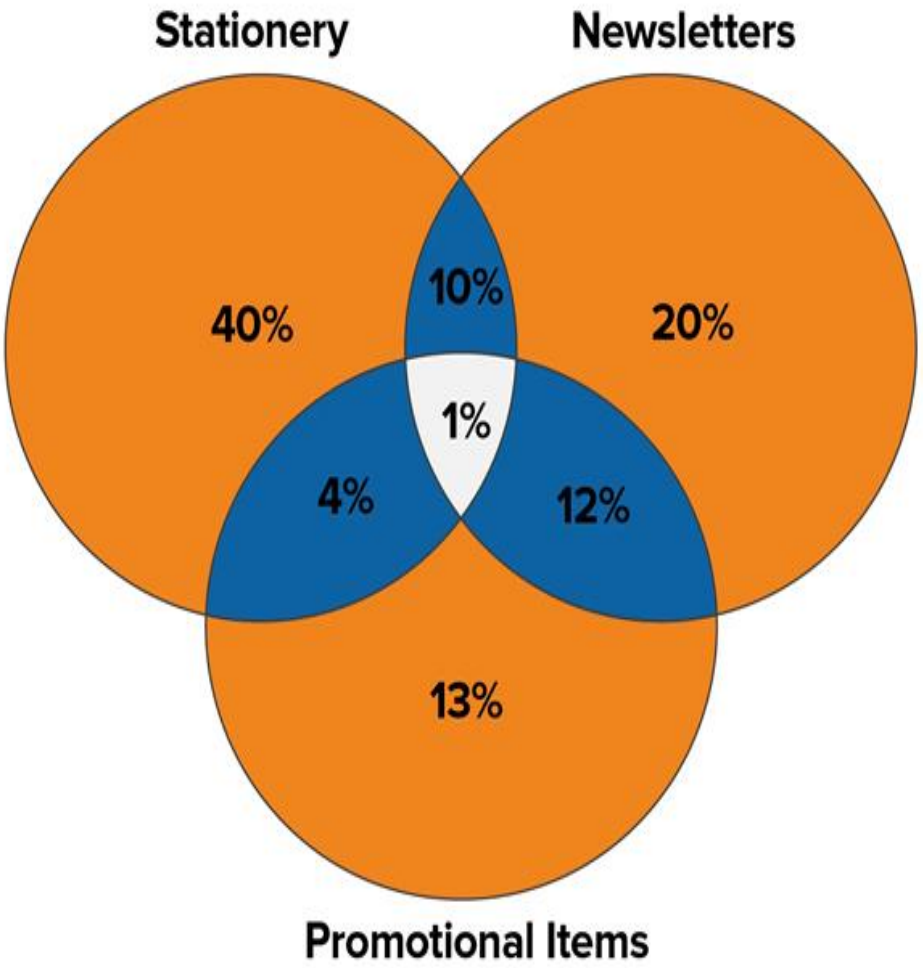
# D-Spot Map

- A map showing the geographic location of people with a specific attribute e.g. cases of an infectious disease
- One dot can depict a fixed number of cases



# Venn Diagrams

- Venn diagrams show the overlaps between sets data.
- Each set is represented by circle. The degree of overla between the sets is depicte by the amount of overlap between the circles.
- A Venn diagram is a good choice when you want to convey either the commor factors or the differences between



# 3-Parameter

**A- Proportion:**

- Some times this proportion is called a ***rate*** if it has a relation to time.

$$\text{Proportion} = \frac{\text{Part}}{\text{Total}} \times 100$$

$$\text{Proportion of Males} = \frac{11}{18} \times 100 = 61\%$$

B- Ratio:

$$\text{Ratio} = \frac{\text{Part a}}{\text{Part b}}$$

$$\text{Male to Female Ratio} = \frac{11}{8} \cong 1.5$$

# Quantitative data presentation:

- .

# Quantitative data presentation

- The age of the students will be used as an example

| Age | Frequency |
|---|---|
| 25 | 1 |
| 28 | 1 |
| 30 | 2 |
| 32 | 1 |
| 33 | 1 |
| 34 | 1 |
| 35 | 1 |
| 36 | 1 |
| 37 | 1 |
| 40 | 2 |
| 41 | 1 |
| 42 | 1 |
| 44 | 1 |
| 45 | 2 |
| 55 | 1 |
| Total | 18 |

| Age group | Frequency |
|---|---|
| 25 | 2 |
| 30 | 5 |
| 35 | 3 |
| 40 | 5 |
| 45 | 2 |
| 50 | 0 |
| 55 | 1 |
| Total | 18 |

| Age group | Frequency |
|---|---|
| 22.5 | 1 |
| 27.5 | 4 |
| 32.5 | 5 |
| 37.5 | 4 |
| 42.5 | 3 |
| 47.5 + | 1 |
| Total | 18 |

***It is important to remember that frequency tables show the following:***

***1- The commonest value (of age).***

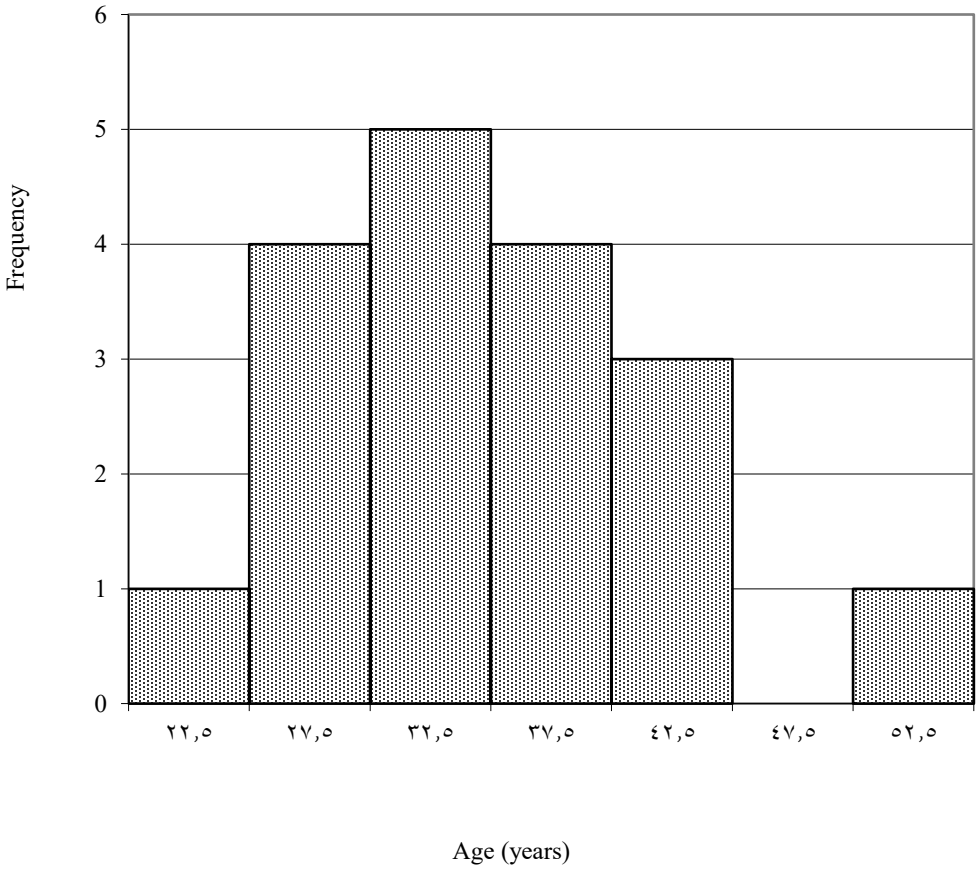
***2- The smallest as well as the largest values or groups.***

***3- The presence of extremely larger or smaller values.***

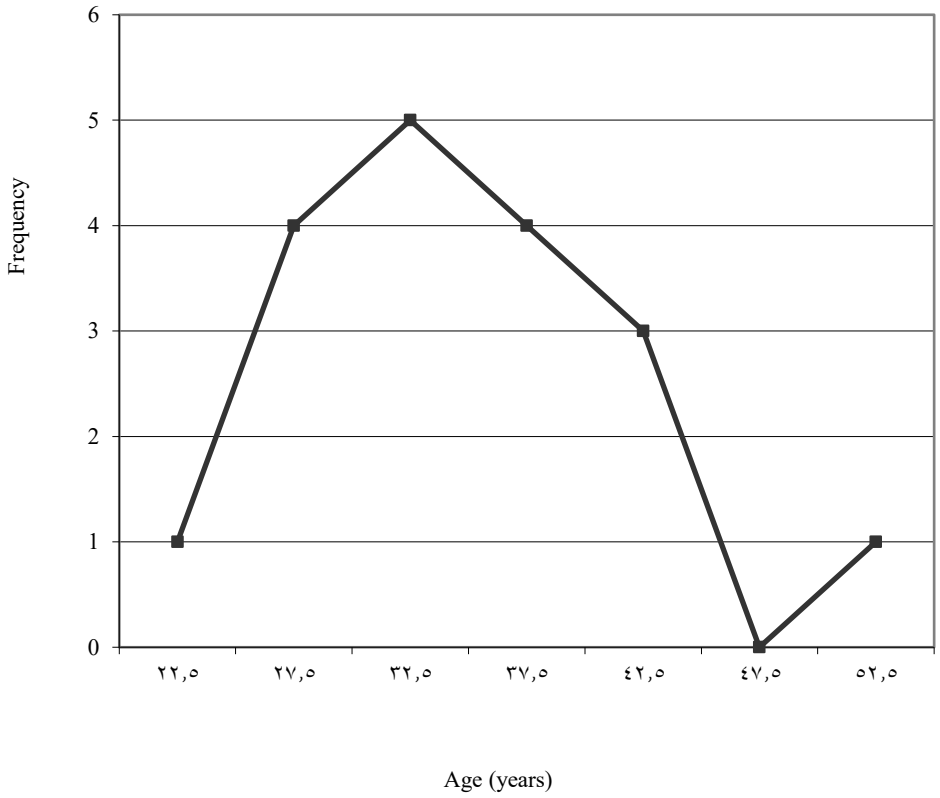
# Graphic presentation

**A: HISTOGRAM**: This is similar to the bar chart of qualitative data that was used to represent sex distribution. The difference is that the columns in the histogram is adjacent to each other, while in bar chart it is separate. This should be clear as in the bar chart each column show different group, while histogram show a continuous variable with the beginning of each group immediate to the end of the preceding one.

A

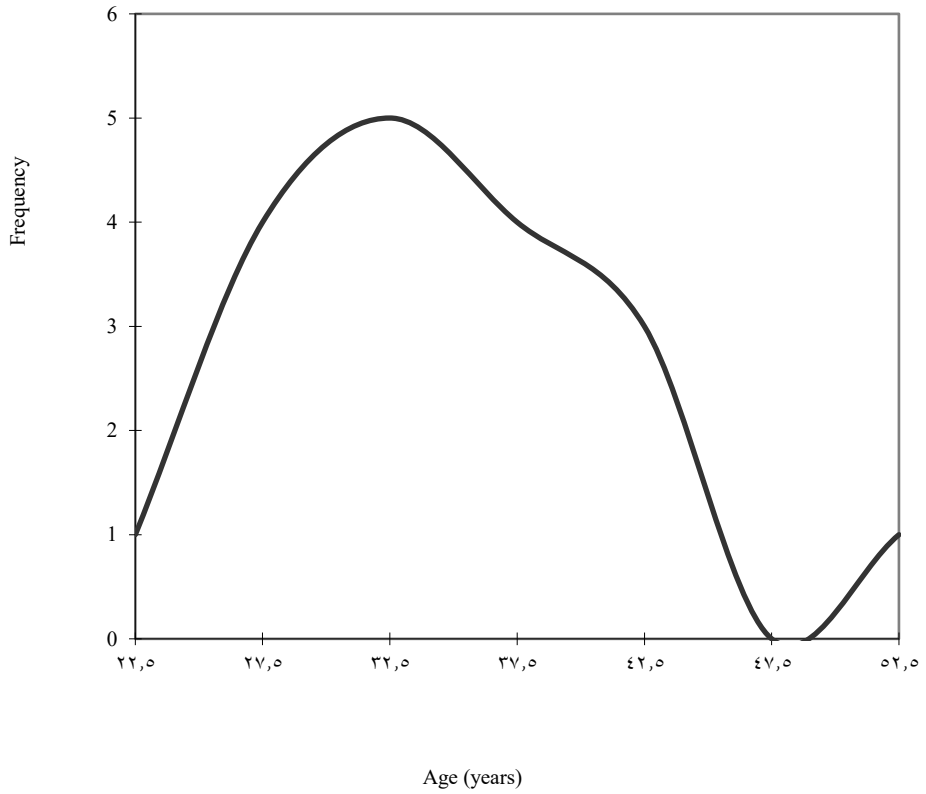


B: FREQUENCY POLYGON: Instead of filling the whole figure with columns, a point is taken at the middle and top of each column. A line is drawn to link each two points together and accordingly multiple lines are seen, hence the name polygon (many lines)



- C: FREQUENCY CURVE: The smooth line that passes between the points instead of the polygon is known as curve.

C



# Parameters

These two measures present the following:

- 1- The center, middle or common value. (central tendency measure)
- 2- The homogeneity or heterogeneity of the values. (Dispersion measure)

# Central tendency measure

- **Mode** The most frequent value or the value that was seen more than the others.

| Grade | 40 | 45 | **50** | 55 | 60 | 66 | 67 | 70 | 72 | 75 | 80 | 90 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1 | 1 | **3** | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 |

- **Mid Range** The value that lies midway between the highest (maximum) and lowest (minimum) values.

$$\text{Mid range} = \frac{\text{Maximum} + \text{Minimum}}{2}$$

- **Median:** The value that belongs to the individual in the middle of an arranged group of values from the lowest to the highest (ascending).

To find the median:

- Arrange the data points from smallest to largest.
- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list.

## Example 1

- Find the median of this data:

111, 444, 222, 555, 000

1-Put the data in order first:

000, 111, 222, 444, 555

2-There is an odd number of data points, so the median is the middle data point.

000, 111, **222**, 444, 555

- **The median is 222**.

Example 2

- Find the median of this data:

10,40,20,50

1- Put the data in order first:

10,20,40,50

2-median=$\frac{20+40}{2}$ =30

- (Average) (Mean) The value that could be given to each individual, when the total values are shared equally.

$$\text{A. mean} = \frac{\text{Sum of values}}{\text{Number of individual s}} = \frac{\sum x}{n} = \overline{X}$$

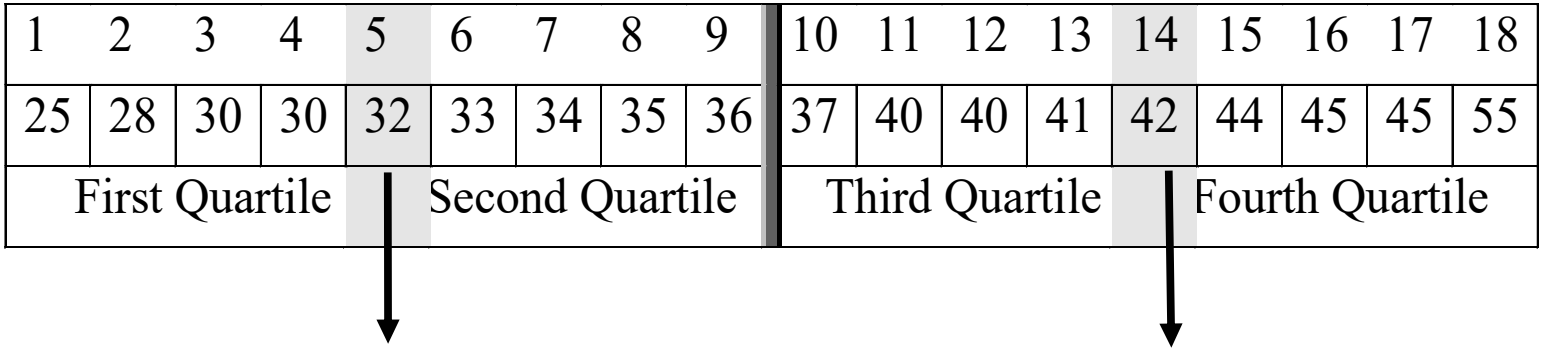
$$\text{A. mean} = \frac{25+28+30+30+..+55}{18} = \frac{672}{18} = 37.3 years$$

# 2- Measures of DISPERSION:

- Minimum and maximum: The highest value and the lowest value. In this example it is 25 and 55 years

- Range: The difference between the highest value and the lowest value. The range is (55 – 25 = 30 years)

- **Quartiles and Percentiles**: The age range of this group of 18 students is 55-25=30 years. If the older student was not present then the range would have been 45-25=20 years. This means that a single value could give a non-real wide range of the group's age. Since we can not ignore a single value, and also we do not want to give a wrong impression we estimate the interquartile range.



- The values are arranged in an ascending order i.e. from the lowest to the highest as shown in figure 11. The group is then divided into four equal parts, each part contains one quarter of the observations. ***In this example each quarter would contain 18 / 4 = 4.5 individuals.*** The value of the ***$5^{th}$ individual*** is the minimal value of the interquartile range. The value of the student number 14 (18/4 x 3 = 13.5) is the maximum of the interquartile range. The interquartile range of these students is 42-32=10 years.

- Percentiles is used when the number of observations is large. Also, arrange individuals according to their values form the lowest to the highest. When the individuals are 100 then the lowest value would be 1st percentile or centile and the highest is 100th centile.

- **Standard deviation (SD)**: It expresses the mean differences (deviations) around the mean value.

$$SD = \sqrt{\frac{\sum \left(x - \overline{x}\right)^2}{n - 1}}$$

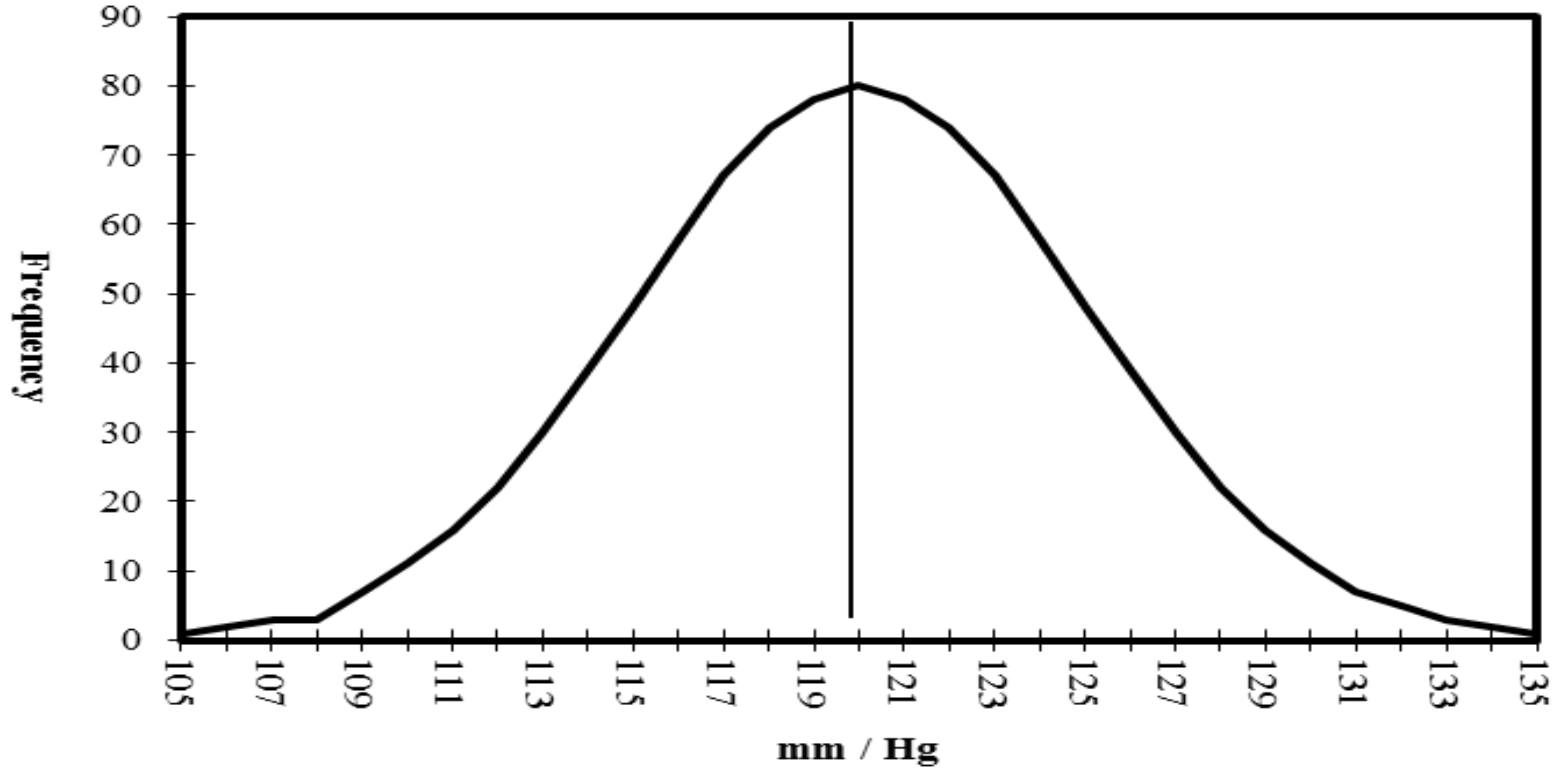
- step 1: Find the mean.
- Step 2: For each data point, find the square of its distance to the mean.
- Step 3: Sum the values from Step 2.
- Step 4: Divide by the number of data points.
- Step 5: Take the square root.

| | Age | Column 2 | Column 3 | Column 4 | Column 5 |
|---|---|---|---|---|---|
| | **X** | X-Ẋ | | (X-Ẋ)2 | |
| | 25 | -12.333 | 12.3 | 152.1 | 625 |
| | 28 | -9.333 | 9.3 | 87.1 | 784 |
| | 30 | -7.333 | 7.3 | 53.8 | 900 |
| | 30 | -7.333 | 7.3 | 53.8 | 900 |
| | 32 | -5.333 | 5.3 | 28.4 | 1024 |
| | 33 | -4.333 | 4.3 | 18.8 | 1089 |
| | 34 | -3.333 | 3.3 | 11.1 | 1156 |
| | 35 | -2.333 | 2.3 | 5.4 | 1225 |
| | 36 | -1.333 | 1.3 | 1.8 | 1296 |
| | 37 | -0.333 | 0.3 | 0.1 | 1369 |
| | 40 | 2.667 | 2.7 | 7.1 | 1600 |
| | 40 | 2.667 | 2.7 | 7.1 | 1600 |
| | 41 | 3.667 | 3.7 | 13.4 | 1681 |
| | 42 | 4.667 | 4.7 | 21.8 | 1764 |
| | 44 | 6.667 | 6.7 | 44.4 | 1936 |
| | 45 | 7.667 | 7.7 | 58.8 | 2025 |
| | 45 | 7.667 | 7.7 | 58.8 | 2025 |
| | 55 | 17.667 | 17.7 | 312.1 | 3025 |
| **Sum** | 672 | 0.0 | 106.7 | 936.0 | 26024 |
| **A. Mean** | **37.333** | | | | |
| **SD** | | | | **7.42** | |

$$SD = \sqrt{\frac{26024 - \frac{(672)^2}{18}}{17}} = \sqrt{\frac{26024 - \frac{451584}{18}}{17}} = \sqrt{\frac{26024 - 25088}{17}} = \sqrt{\frac{936}{17}} = \sqrt{55.06} = \pm 7.42$$

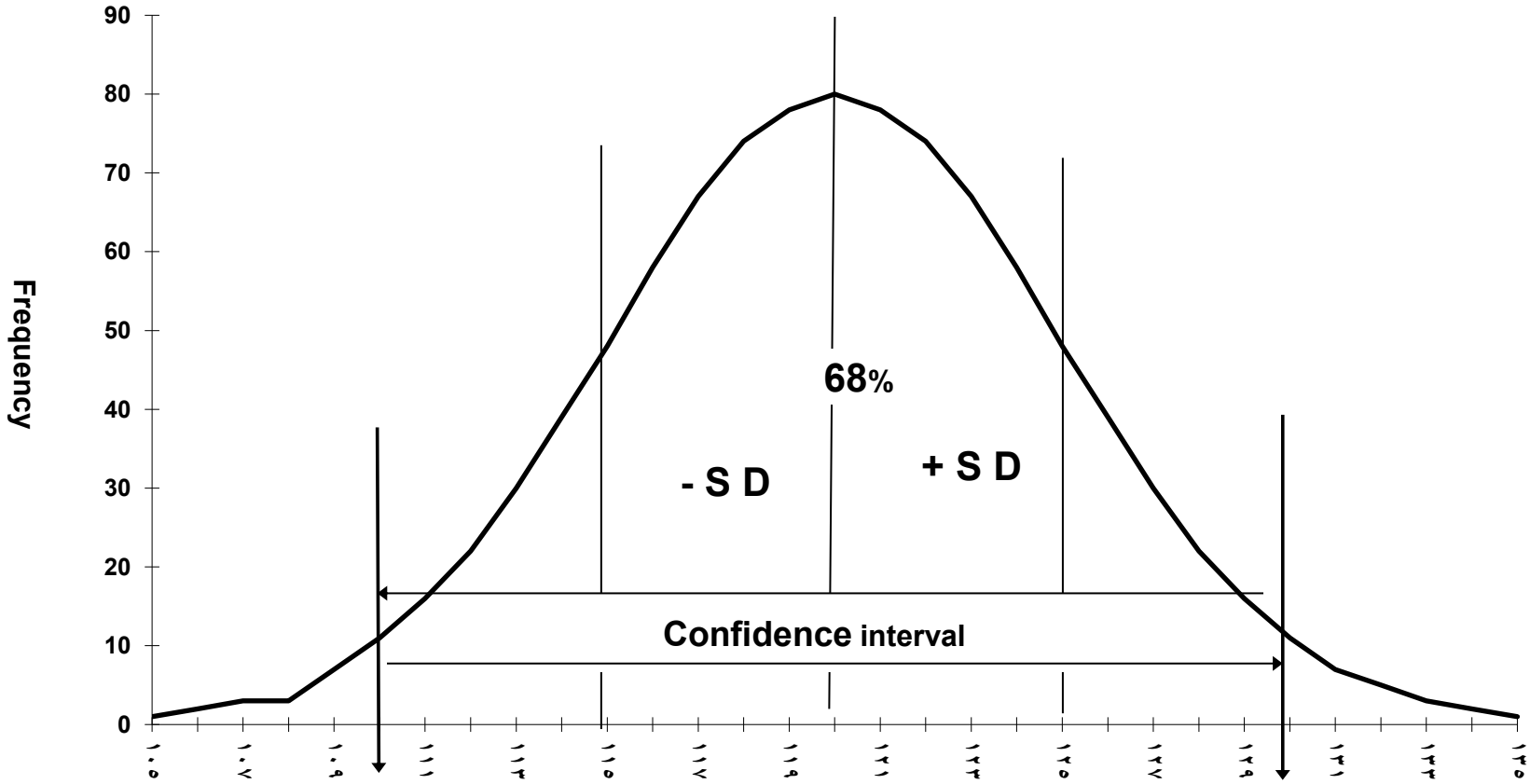
# NORMAL DISTRIBUTION CURVE

Example: A group of 1000 adult males systolic blood pressure as measured in mm / Hg is presented in the following curve:



- This is known as the normal distribution curve. It is ***normal as regards the distribution*** and does not mean it shows normal blood pressure values.
- It has the following characteristics:

1. It has a peak and two symmetrical sides.
2. The peak coincides with all measures of central tendency (mode, mid-range, mean and median).
3. The curve extends to infinity on both sides.




# Areas under the normal distribution curve, and confidence limits

- It is therefore, recommended by most statisticians that ***95% of the individuals in the group is a reasonable proportion that could be considered similar***. These similar 95% could use the mean value as a reasonable approximation to their real values. ***The remaining 5% are too far from the mean and can not use the mean to represent their values.*** Accordingly, ***2.5% in the group have unacceptably high values and another 2.5% have unacceptable low values***.

- we can accept ***differences in values between individuals so long as they are among the closest 95% to the mean value***. On the other hand, ***we do not accept or reject the differences of values that belong to the 5% of individuals furthest to mean***. In other words, we are confident that the mean could represent the closer 95% individual values. The other 5% values are so different for the mean to express and is called significantly different.

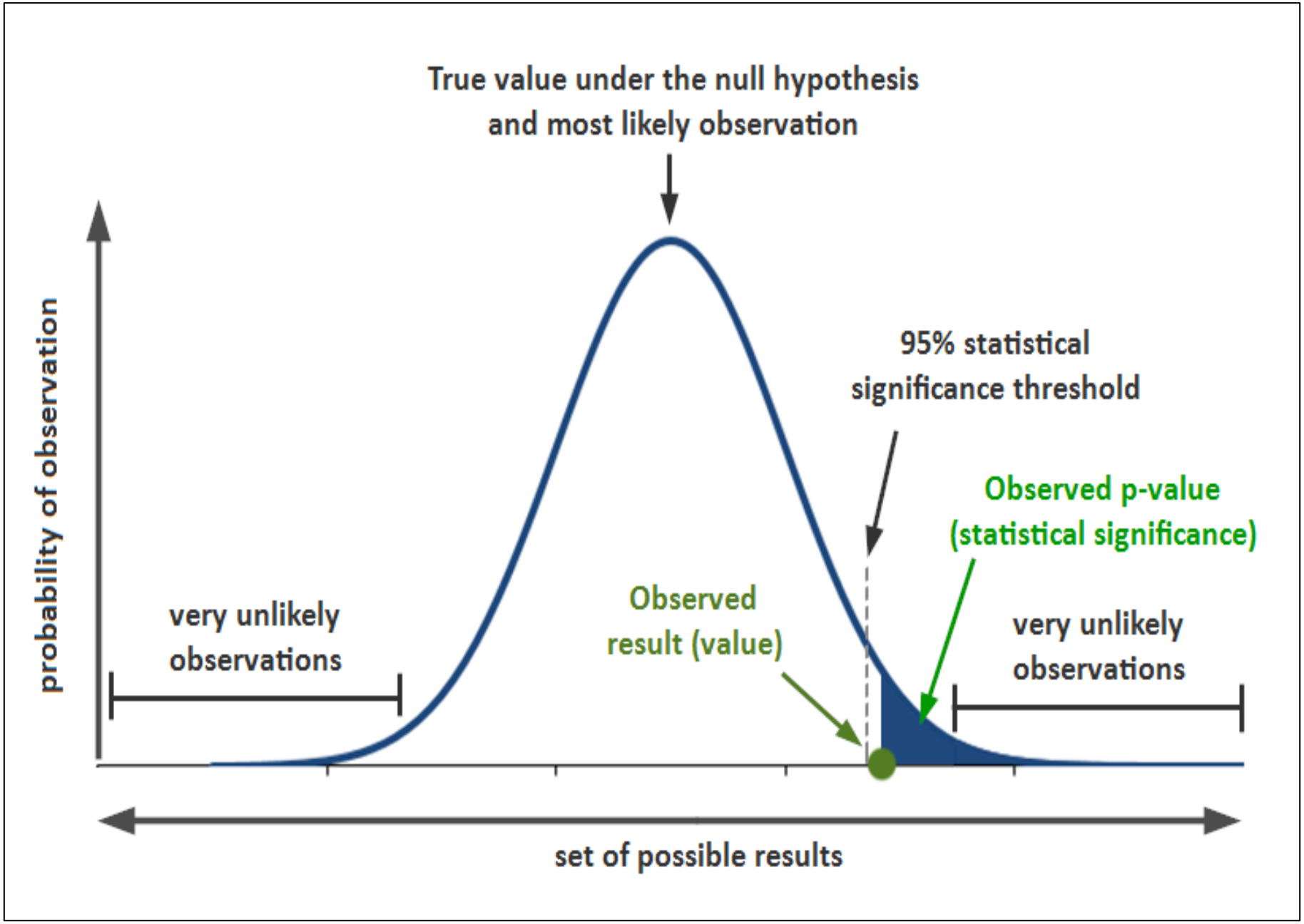
# CONFIDENCE LIMITS

- These are the two values that limit the 95% closer values to the mean. Between these two values we are confident that any value could be represented by the mean and that it belongs to the group. Any value that is more than the high confidence limit is considered significantly different from the mean value, and hence different from the other individuals in the group. The same applies to any value less than the lower confidence limit.

# STATISTICAL SIGNIFICANCE

- “Statistical significance helps quantify whether a result is likely due to chance or to some factor of interest,”

- When a finding is significant, it simply means you can feel confident that’s it real, not that you just got lucky (or unlucky) in choosing the sample.


True value under the null hypothesis
and most likely observation
probability of observation
95% statistical
significance threshold
Observed p-value
(statistical significance)
very unlikely
observations
Observed
result (value)
very unlikely
observations
set of possible results

# Thank you

- https://www.youtube.com/watch?v=dr1DynUzjq0