**Live**
**Love**
**Biostatistics**
**Astaghfurlla bas-.-**

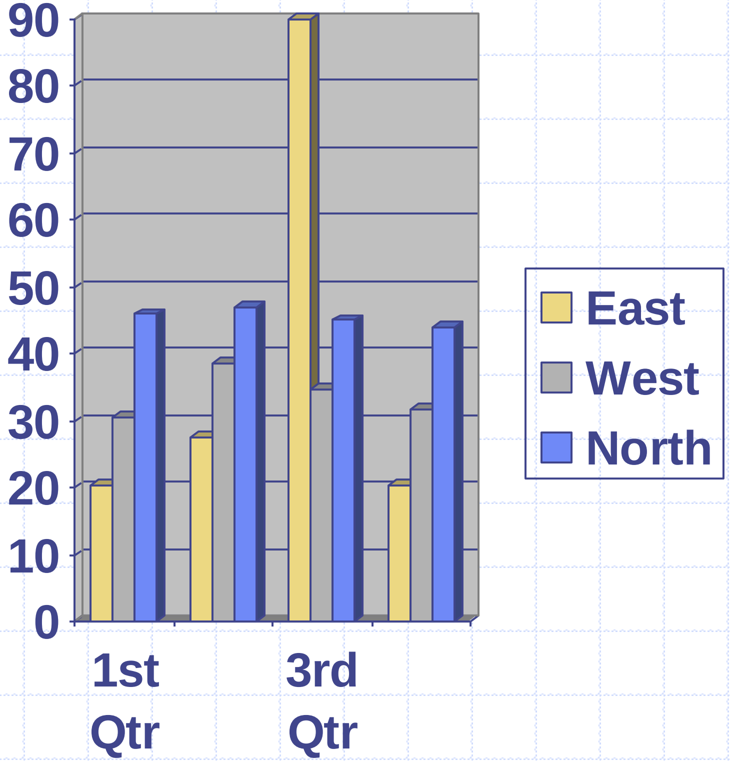
# Descriptive measures

Edited by : D3ana Rida :")

Senpai will never notice you

What is written in red is the added info mentioned by the doctor or just to clarify things , you can refer to the book from page 34-47 (just read them for extra examples, don't go into much details just study the things discussed in slides because in the book there is so much info)

- Capture the main 4 basic Ch.Ch. of the sample distribution:




- Central tendency
- Variability (variance)
- Skewness
- kurtosis

# Measures of central tendency

# MEAN (average)

Mean is affected by extreme values .. Like if I want to measure the mean for marks of students and they were all above 80 expect one is 10 even if it is an outlier , this will drag down the mean with it . So when comparing this mean to another group (has a normal mean, no extreme value) it will not be comparable , our decision will be in valid.

- $M = \sum X/N$
- X: sum of all values
- N: number of values
- The data compatible with measuring mean is at the continuous level (ratio to be more specific)

Note : Nominal level is used with and frequencies percentages

- It is the best average for symmetrical frequency distributions that have a single peak, (normal distribution).

Book (pg 34 -35 )

# Example

| Grades | Frequency |
|---|---|
| 70 | 20 |
| 80 | 50 |
| 50 | 10 |

M= (70*20+80*50+50*10)/ (20+50+10)
So N= 80 not 3 , N is the sum of frequency here not the number of values
But if the data was 70 80 90 45 100 then the mean just adding them and diving them by N=5

# Measures of central tendency MEAN (Ch.Ch of the mean)

1. The sum of deviations of the values from the mean always = Zero.

| X | X-M | $(X-M)^2$ |
|---|---|---|
| 4 | $4 - 6 = -2$ | $(-2)^2 = 4$ |
| 4 | $4 - 6 = -2$ | $(-2)^2 = 4$ |
| 10 | $10 - 6 = 4$ | $(4)^2 = 16$ |
| 5 | $5 - 6 = -1$ | $(-1)^2 = 1$ |
| 7 | $7 - 6 = 1$ | $(1)^2 = 1$ |
| $\sum X = 30$ | $\sum (X - M) = 0$ | $\sum (X - M)^2 = 26$ |
| $N = 5$ | | |
| M (μ) (used when describing the mean for the whole population) = 6 | | |

2. $\sum (X - M)^2$ (THE SUM OF SQUARES) is smaller than the sum of squares around any other value. (least squares).(doctor didn't explain it)
3. A mean of total group (M total = M1n1 + M2n2 + ……...) calculate the mean of the means, I can have 4 groups that I calculated their mean, and get the mean of the 4 means.

- Mean is intended mainly for interval and ratio variables and some times in ordinal variables, but not in nominal ones such as the mean of gender = 0.75.

# Measures of central tendency Median

Book pg 36

- The middle value of a set of ordered numbers
- Also known as $50^{th}$. Percentile (P50)
- Example : p50=60 ,, it means 50% of students have marks less than 60

# Measures of central tendency Median

- The median is not sensitive to extreme scores (e.g. 8, 10, 10, 18, 24, 29, 36, 48, 60, 224) (in the exam they will not be ordered so you need to reorder them ascending or descending)
- Used in symmetrical(even numbers of observations , choose the 2 numbers in the middle and get their average) and asymmetrical distributions (odd numbers of observations, simply choose the number in the middle)

- In the previous example they are even observations , so 24+ 29/2 the result is the median
- Even if there is a repeated observation we don't delete it out, instead count it in
- Another example (numbers mean nothing it's just for explaining)

| | Frequency |
|---|---|
| 70 | 1 |
| 80 | 2 |
| 20 | 3 |
| 50 | 3 |

Start ordering data and writing them without frequencies
20 20 20 50 50 50 70 80 80
The result is 50 since they are odd observations
Median = 50

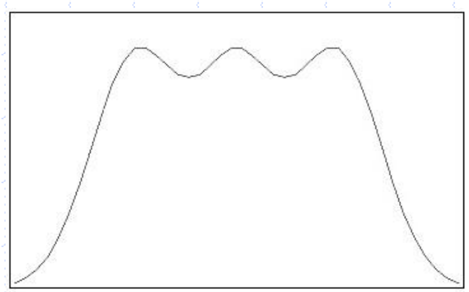
# Measures of central tendency Median

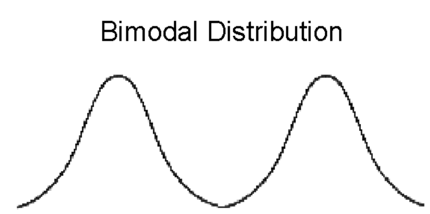
- It is useful when the data are skewed
- Appropriate in ratio, interval and ordinal variables, but not for nominal data.

# Measures of central tendency Mode

Book pg 36-37

- The most frequent value or category in a distribution
- Not calculated, but spotted
- E.g. 8, 10, 10, 18, 24, 36, 48, 60 the mode is 10
- There can be more than one mode, bimodal (2 modes) or trimodal and so on.
- It is appropriate for all variables including the nominal ones.





- Studies that use mean considered parametric

# Comparison of Central Tendency Measures

Book pg 37-40

In a perfect world, the mean, median & mode would be the same.

However, the world is not perfect & very often, the mean, median and mode are not the same,

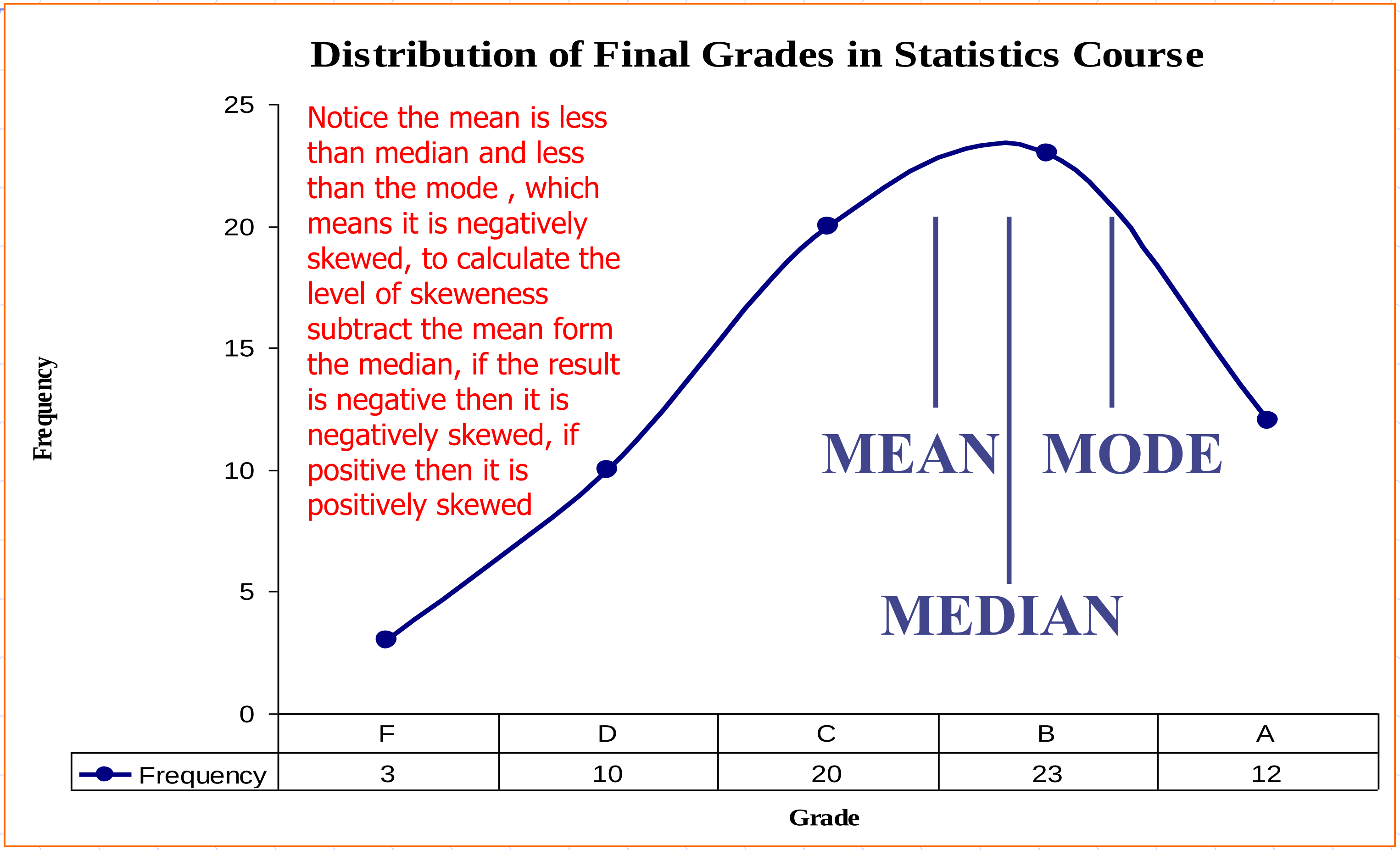
# Summary for central tendency measures

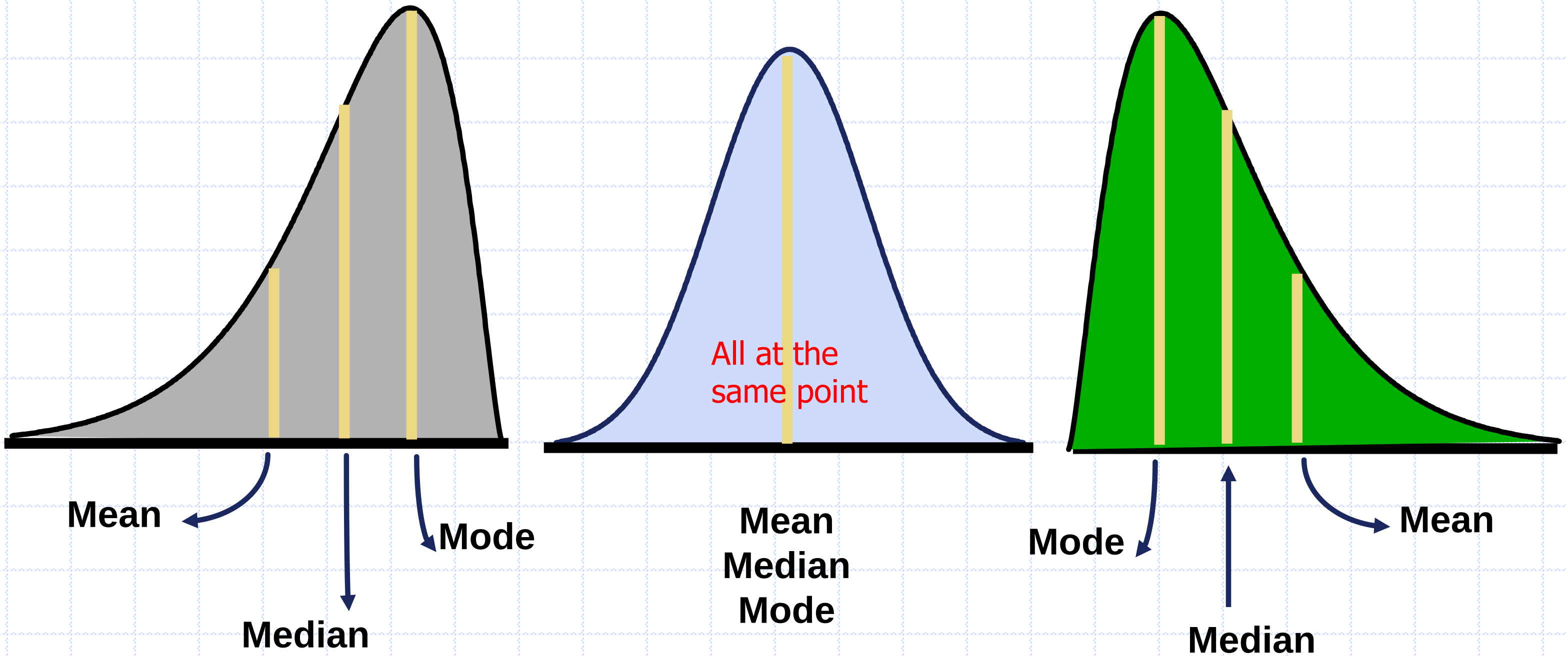
- Use mean as more frequent unless the distribution is badly skewed (when the data is skewed we compare by the median, because it is not affected by extreme values)

Use mode for nominal variables

If the mean is greater than median, the distribution is positively skewed.

# Central Tendency - Graphed






**Negatively Skewed**

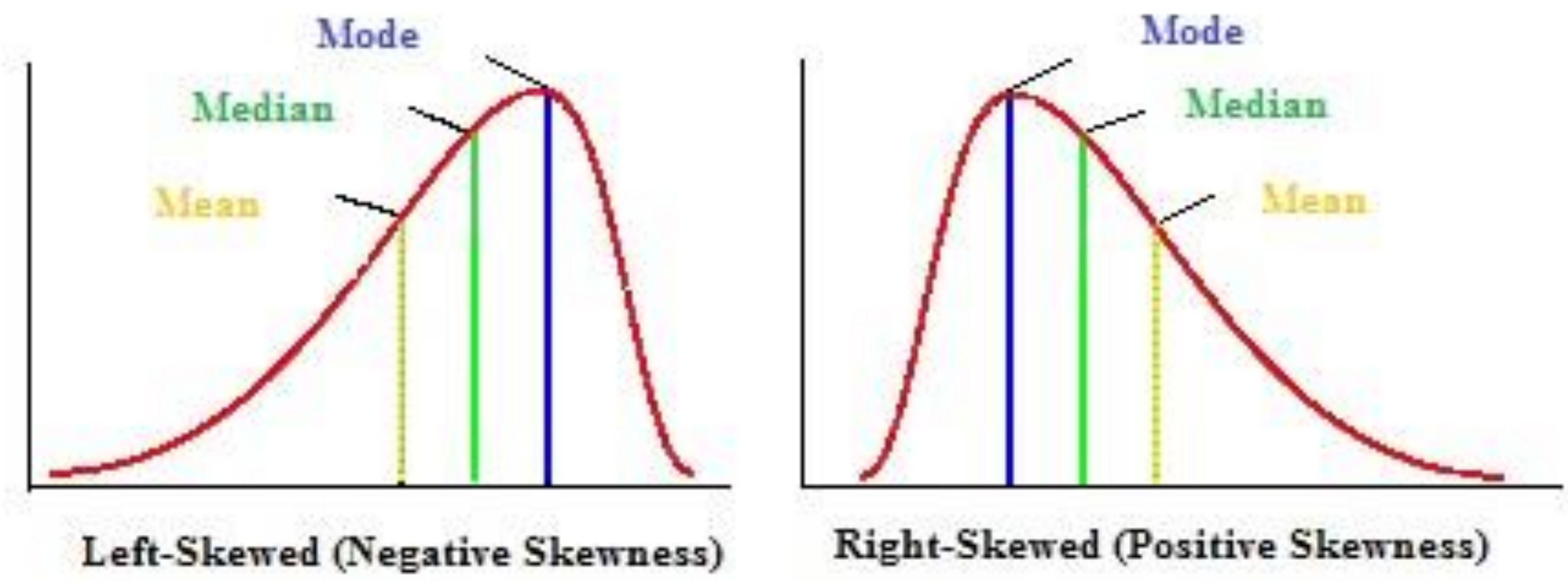
**Symmetric (Not Skewed)**

Perfect normal distribution
bell shaped

**Positively Skewed**

Mean is smaller than the median (x axis)

The mean is bigger than the median (x axis)




If the data is severely skewed we use non parametric statistics , we will get to that later

# Comparison of Central Tendency Measures

- **Use Mean** when distribution is reasonably symmetrical, with few extreme scores and has one mode.
- **Use Median** with nonsymmetrical distributions because it is not sensitive to skewness.
- **Use Mode** when dealing with frequency distribution for nominal data

# Measures of variability, scatter or dispersion (SD) standard deviation

Book pg 41-44 things not discussed here don't study them

- SD (in capital letters is used when talking about the whole population) = square root of $\sum (X - M)^2 / n - 1$
- The equation above we don't have to memorize it.
- Every value in the distribution entered in calculation of SD.
- SD is a measure of variability around the mean (how far my observation is far from the mean, whenever reporting the mean you need to report with it the standard deviation).
- It is sensitive to extreme values
- It serves best in normally distributed populations
- It shows how much the data is scattered and high variability, which means I have a lot of confounding factors affecting my results

# Measures of variability, scatter or dispersion (Range)

Book pg 44

- The difference b/w the maximum and the minimum values in a distribution

- Sensitive to extreme values
- The higher the range the more variable is the data, which makes a lot of external variables to effect on the relationship between independent and dependent variables

# Measures of variability, scatter or dispersion (percentile) will be discussed in more details in another lecture

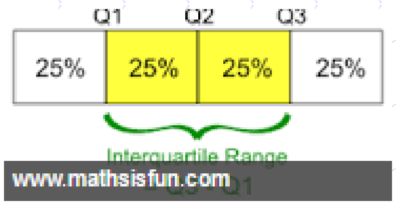
- Is a score value above which and below which a certain percentage of values in a distribution fall.

- *P60 = 30 means that 60% of the values in the distribution fall below the score 30.*

# Measures of variability, scatter or dispersion (percentile)

- It allows to describe a score in relation to other scores in the distribution.
- $25^{th}$. percentile = first quartile
- $50^{th}$.percentile = second quartile (median)
- $75^{th}$. percentile= third quartile

The **interquartile range** (**IQR**) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.



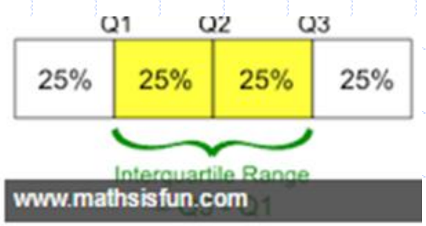
# Comparison of Measures of Variability

Book pg 44 -45

**Interpercentile Measures** **The more interquartile range the more variable it is.**

- Easy to understand
- Can be used with distributions of any shape
- Especially useful in very skewed distributions
- Use IQR when reporting median of distribution

The **interquartile range** (**IQR**) is a measure of variability, based on dividing a data set into quartiles. Quartiles divide a rank-ordered data set into four equal parts. The values that divide each part are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively.



# Comparison of Measures of Variability

Book pg 45 - 44

**Standard Deviation** **(cant compare 2 groups when they have different SD)**

- Most widely used measure of variability
- Most reliable estimate of population variability
- Best with symmetrical distributions with only one mode

# Comparison of Measures of Variability

**Range**

- Main use is to call attention to the two extreme values of a distribution
- Quick, rough estimate of variability
- Greatly influenced by sample size: the larger the sample, the larger the range

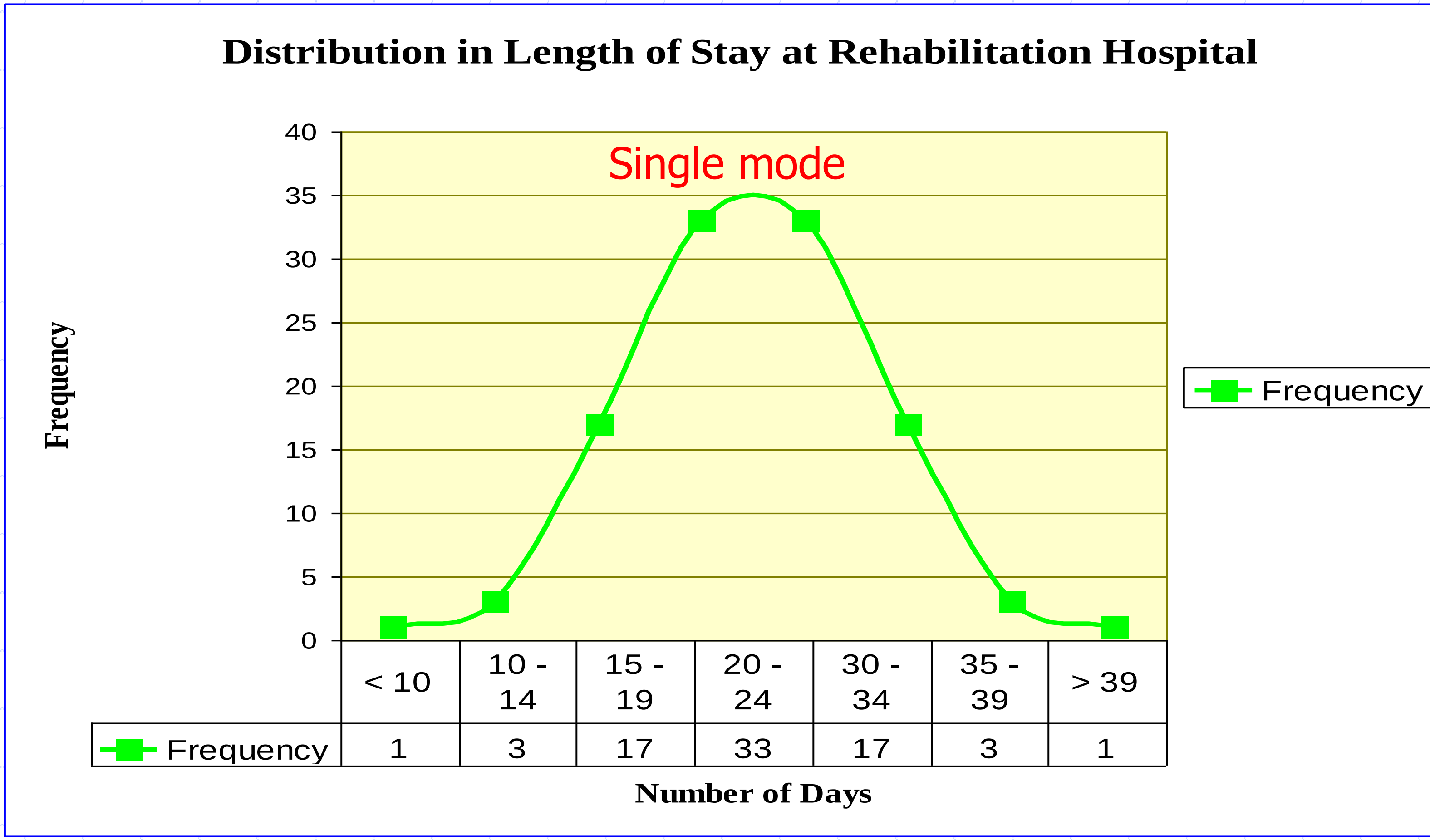
# Summary of variability measures

- SD the most frequently used measure (normal curve = one mode)
- Range is a rough estimate of variability (influenced by sample size)
- Range and percentiles are useful in skewed distributions.
- There are no measures of variability for nominal variables.

# Shape of the Distribution

- The shape of the distribution provides information about the central tendency and variability of measurements.
- Three common shapes of distributions are:
  - Normal: bell-shaped curve; symmetrical
  - Skewed: non-normal; non-symmetrical; can be positively or negatively skewed
  - Multimodal: has more than one peak (mode)

# Normal Distribution

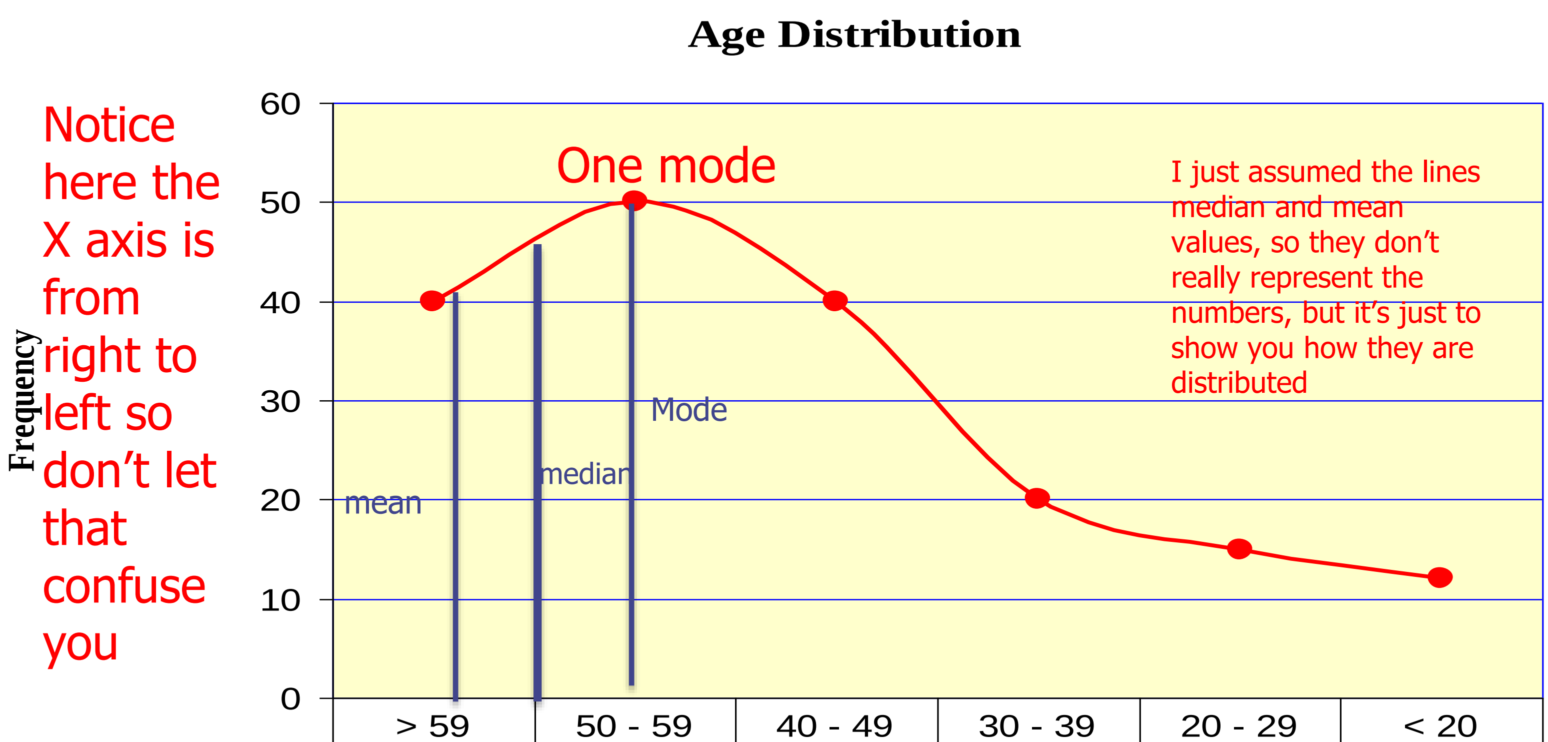



| | < 10 | 10 - 14 | 15 - 19 | 20 - 24 | 30 - 34 | 35 - 39 | > 39 |
|---|---|---|---|---|---|---|---|
| Frequency | 1 | 3 | 17 | 33 | 17 | 3 | 1 |

# Positively Skewed Distribution

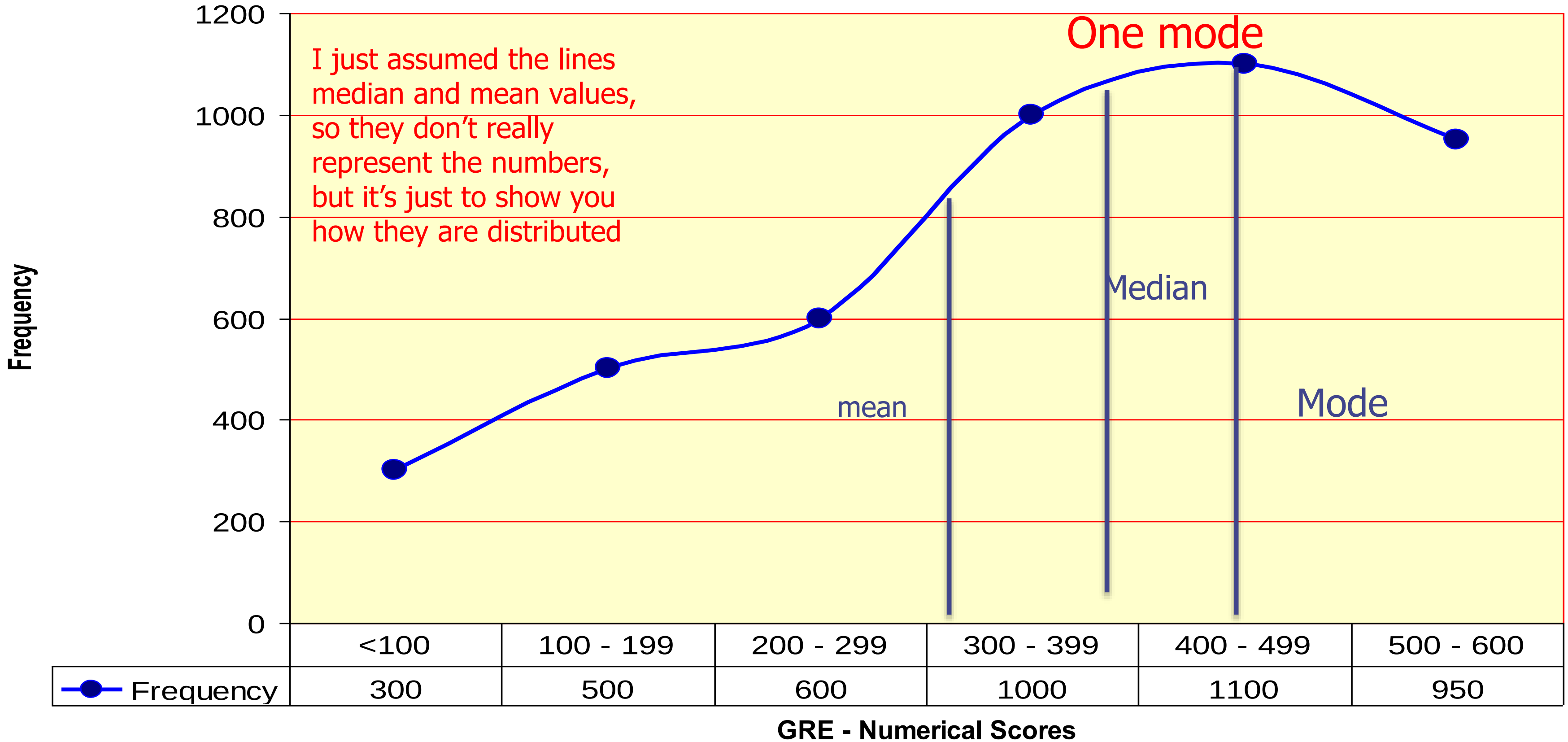
Don’t visually decide on the curve, better to do calculations

Notice here the X axis is from right to left so don’t let that confuse you



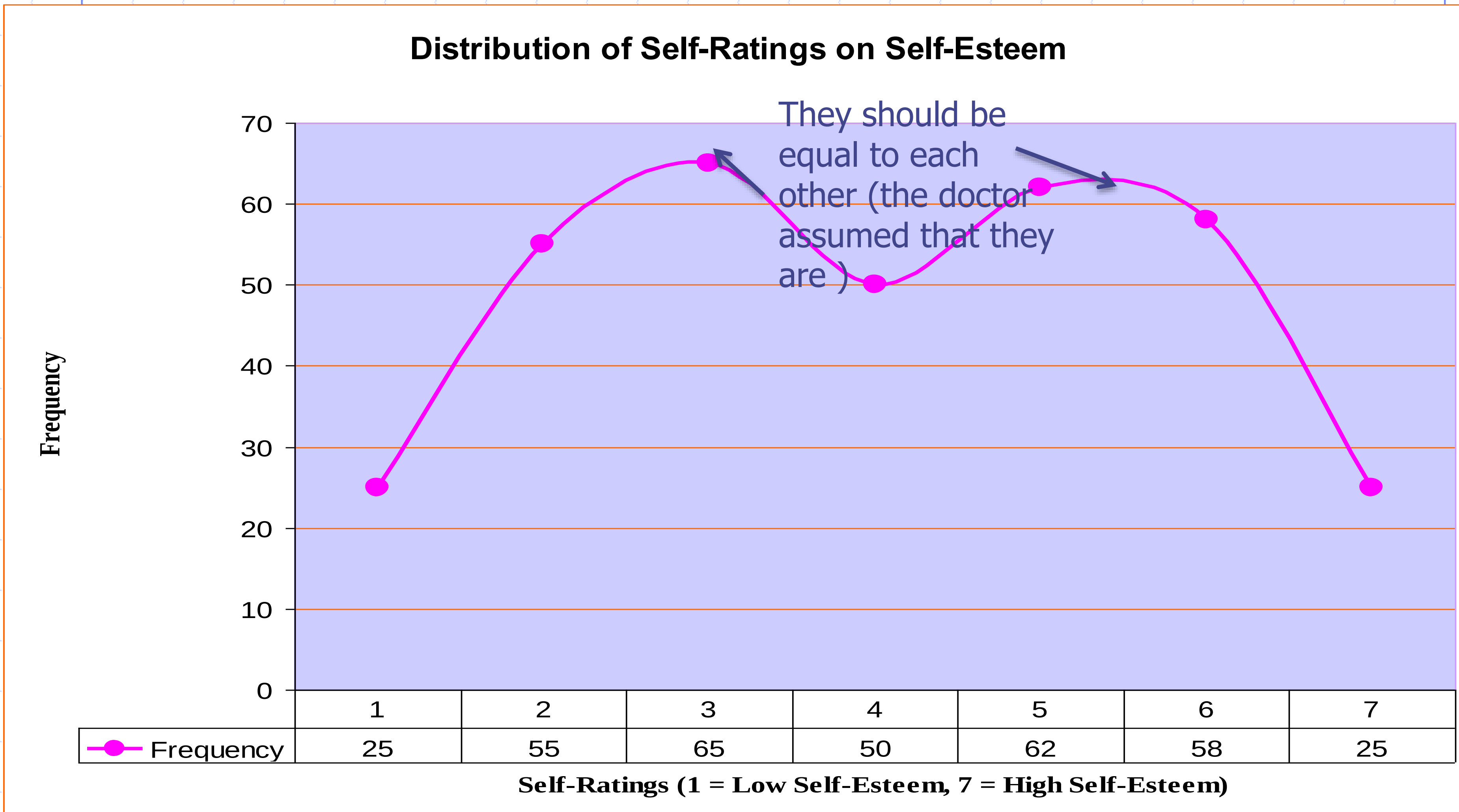

| | > 59 | 50 - 59 | 40 - 49 | 30 - 39 | 20 - 29 | < 20 |
|---|---|---|---|---|---|---|
| Frequency | 40 | 50 | 40 | 20 | 15 | 12 |

# Negatively Skewed Distribution




| | <100 | 100 - 199 | 200 - 299 | 300 - 399 | 400 - 499 | 500 - 600 |
|---|---|---|---|---|---|---|
| Frequency | 300 | 500 | 600 | 1000 | 1100 | 950 |

# Bimodal Distribution



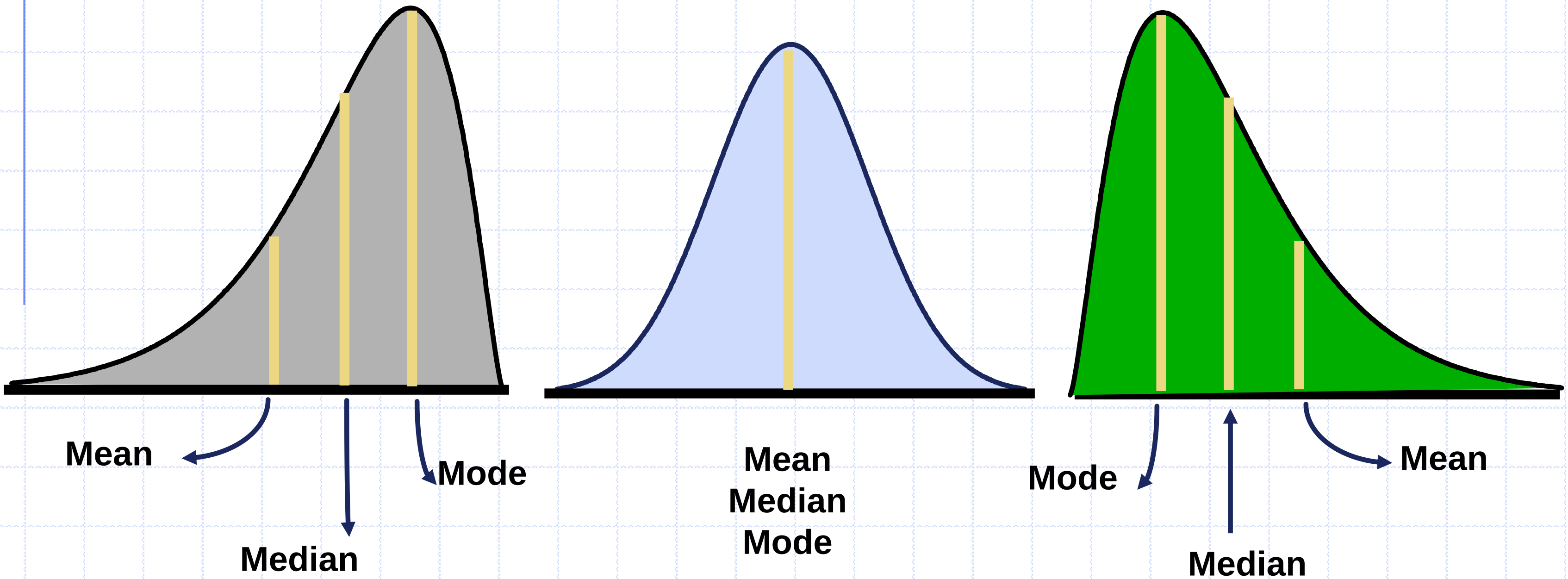
# Variable Distribution Symmetry

- Normal Distribution is symmetrical & bell-shaped; often called “bell-shaped curve”
- When a variable’s distribution is non-symmetrical, it is skewed
- This means that the mean is not in the center of the distribution

# Skewness

- Skewness is the measure of the shape of a nonsymmetrical distribution
- Two sets of data can have the same mean & SD but different skewness
- Two types of skewness:
  - Positive skewness
  - Negative skewness

# Relative Locations for Measures of Central Tendency

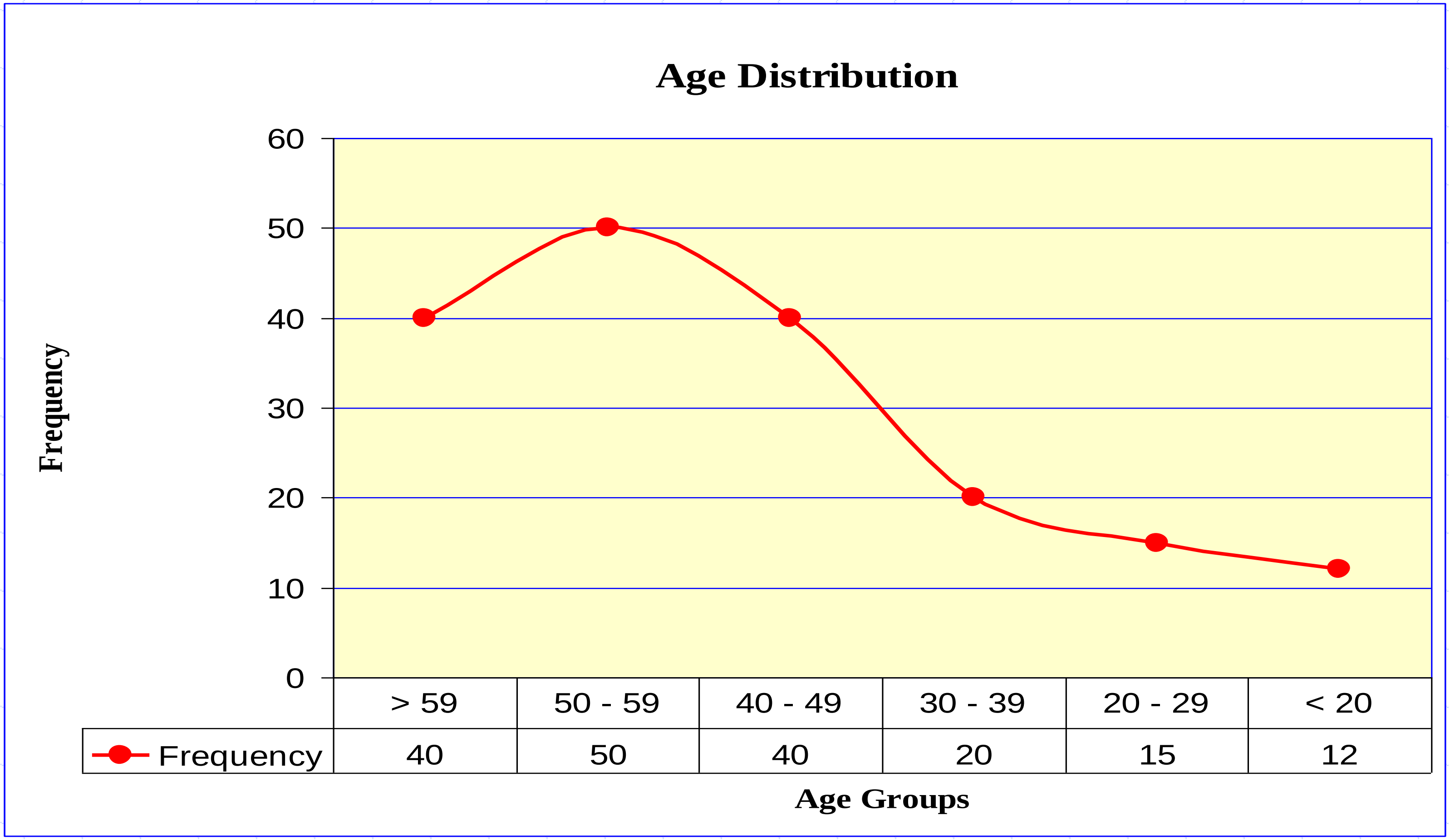



Negatively Skewed

Symmetric (Not Skewed)

Positively Skewed

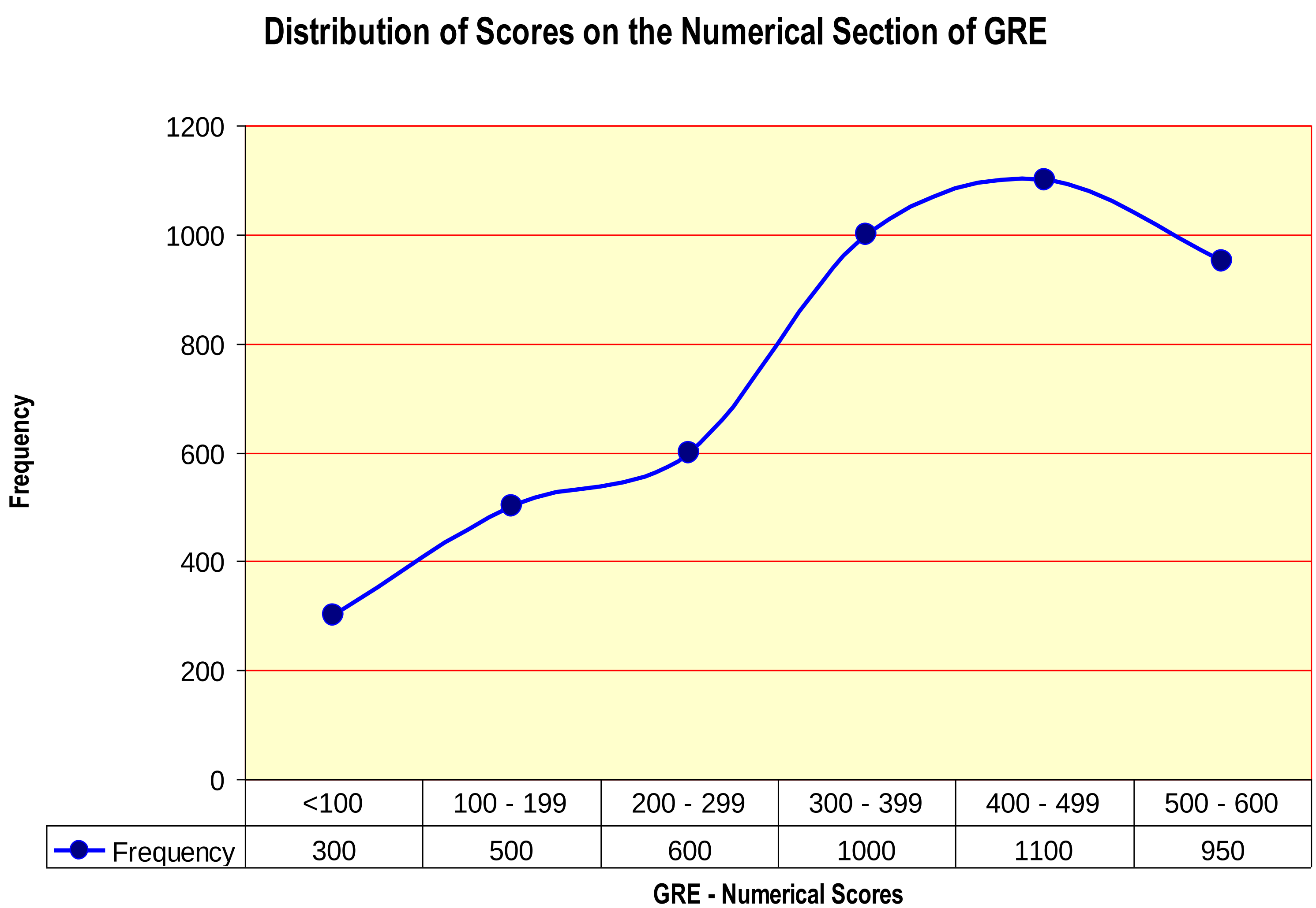
# Positively Skewed Distribution



# Positive Skewness

- **Has pileup of cases to the left & the right tail of distribution is too long**

# Negatively Skewed Distribution




| | <100 | 100 - 199 | 200 - 299 | 300 - 399 | 400 - 499 | 500 - 600 |
|---|---|---|---|---|---|---|
| Frequency | 300 | 500 | 600 | 1000 | 1100 | 950 |

# Negative Skewness

- **Has pileup of cases to the right & the left tail of distribution is too long**

# Measures of Symmetry (4 equations use the suitable one depending on what information you have)

Book pg 46 -48

- **A general equation = mean-median**
- **Pearson's Skewness Coefficient Formula =**

$$\frac{\text{mean-median}}{SD}$$

**0.2** **(cut of point ,where your results must not exceed it to be considered not skewed , only when using pearson )**

- **Skewness values > 0.2 or < 0. 2 indicate severe skewness**
- **If negative then it is negatively skewed visa versa**

# Measures of Symmetry

- Fisher's Skewness Coefficient Formula =

$$\frac{\text{Skewness coefficient}^{NB}(\text{calculated in pearson})}{\text{Standard error of skewness}}$$

- *Skewness values >+1.96 SD* indicate severe skewness
- If negative then it is negatively skewed visa versa (so -2 is not less than 1.96 we forget the negative it only shows direction)

NB: Calculating skewness coefficient & its standard error is an option in most descriptive statistics modules in statistics programs

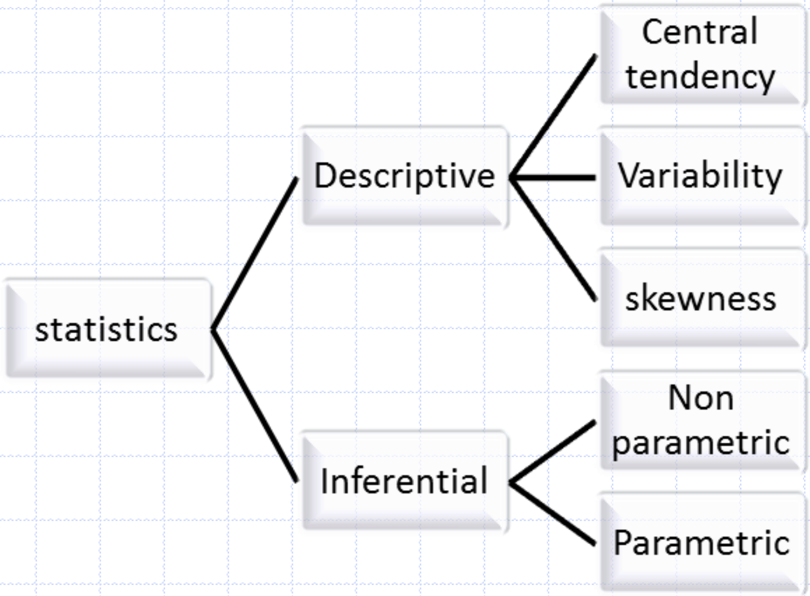
A measure of skewness is Pearson's Coefficient of Skew.

It is defined as:

Pearson's Coefficient = 3(mean - median)/ standard deviation

No cut of point, this equation to show if my data is deviated 3 standard deviations

This will be explained better in another slide




Two important notes :

1- Don't you ever decide or take a decision on data based on descriptive statistics .

2-Descriptive statistics "overview about distributions or data we collect"
if you have two groups "control and experimental" with two means and you were asked to compare between the groups.
for example : we have two diets to reduce the weight, diet A and diet B so, after 3 months of testing them on an experimental and control group, we found that
the experimental group that used diet A reduced the mean of the weights from 120 to 80 Kg , while the control group that used diet B reduced their mean of the weights to 75.
-now we CAN'T say that diet B is better than diet A referring to the difference between the means using "descriptive statistics" the right decision must be taken using inferential statistics
Added by Mohammad da'as

(not included to the end of the slides )

# Data Transformation

- With skewed data, the mean is not a good measure of central tendency because it is sensitive to extreme scores
- May need to transform skewed data to make distribution appear more normal or symmetrical
- Must determine the degree & type of skewness prior to transformation

# Data Transformation

- If positive skewness, can apply either square root (moderate skew) or log transformations (severe skew) directly
- If negative skewness, must "reflect" variable to make the negative skewness a positive skewness, then apply transformations for positive skew

# Data Transformation

- Reflecting a variable change in the meaning of the scores.
  - Ex. If high scores on a self-esteem total score meant high self-esteem before reflection, they now mean low self-esteem after reflection

# Data Transformation (

- As a rule, it is best to transform skewed variables, but keep in mind that transformed variables may be harder to interpret
- Once transformed, always check that transformed variable is normally or nearly normally transformed
- If transformation does not work, may need to dichotomize variable for use in subsequent analyses

# Kurtosis

A measure of whether the curve of a distribution is:

- Bell-shaped -- Mesokurtic
- Peaked -- Leptokurtic
- Flat -- Platykurtic

# Fisher's Measure of Kurtosis

- **Formula = Kurtosis coefficient [NB] / Standard error of kurtosis**
- **Kurtosis values >±1.96 SD indicate severe kurtosis**

**NB: Calculating kurtosis coefficient & its standard error is an option in most descriptive statistics modules in statistics programs**

- Practice exercises on skewness and kurtosis)
- Histograms
- Bar Charts
- Box plots
- Scatter plots
- Line charts