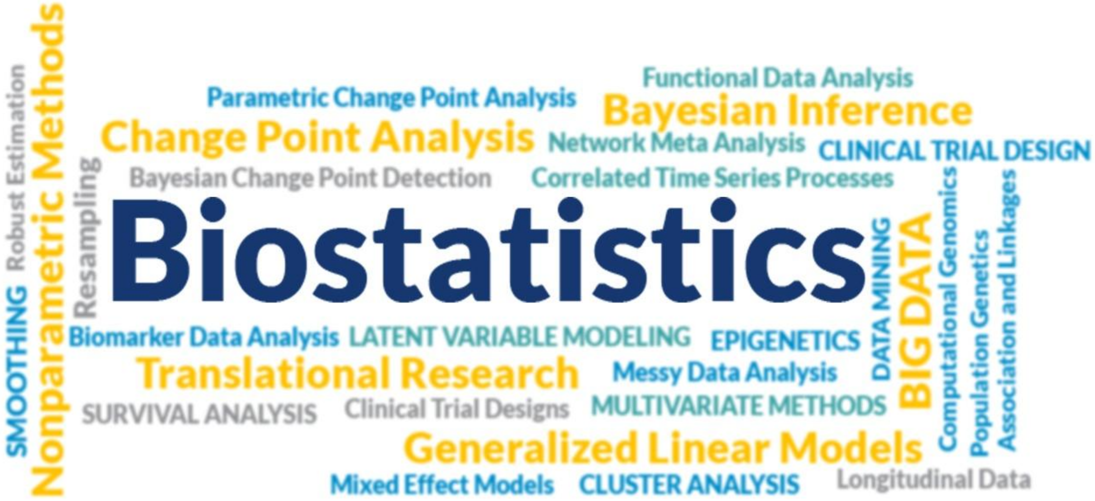


● Sheet ○ Slides

Number

6

Done by:

عبدالله الشنيكات

corrected by:

عبدالله نمر

Doctor

[Note: Study this sheet and the slides simultaneously!]

We will continue with descriptive biostatistics;

The third measure of central tendency that we are going to discuss is the **mode**. The mode is the value that occurs with the greatest frequency.

We may have: no mode, one mode or more than one mode.

in this data set 1,2,3,4,5,6,7,8,9,0 we have no mode. While in this one 0,1,1,2,2,3,3,4,4,5,5,6,6,7 there are 6 modes! **This is in contrary with the mean and median** (there is always one mean and one median)

-Comparison between mean, median and mode:

- The mean and median may not be part of the data set but the mode must be a part of the set if present.
- The mean is sensitive to the extreme values while the other two are not. The mode is not affected by extreme values
- but if there is more than one mode the data isn't normally distributed because it'll have more than one belly "normally distributed data has one belly".

-> if **Mean=Median=Mode** then the data is normally distributed and symmetric.

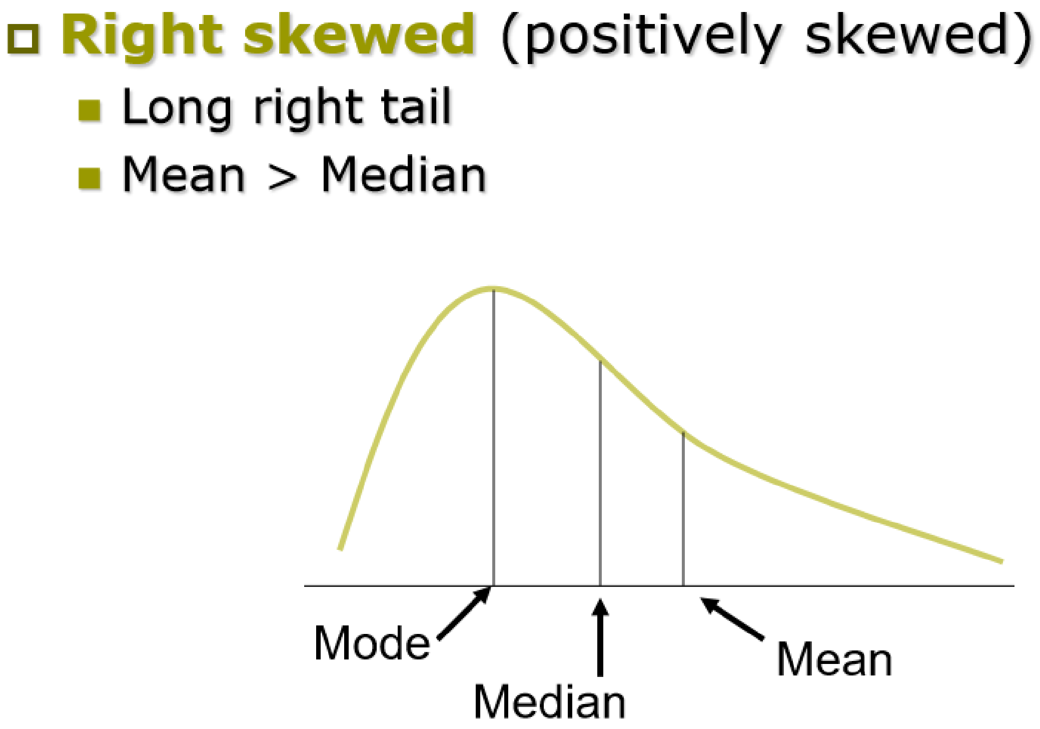
Why do we study this? Who was the first one?

- A French scientist studied traits among people and he found that most people are similar in a certain characteristic. for example, people are around 170cm which is the mean and median and mode (normally distributed), he also studied the standard deviation among them (i.e. how far apart the values from the mean) he noticed that 68% are between 1 standard deviation from the mean [from our example the St.d was 5 so the values are between 165 and 175.] these are characteristics of the normal distribution.

- 1 st.d around the mean -> 68%
- 2 st.d around the mean -> 96%
- 3 st.d around the mean -> 99%
- Remember! We have normal distribution and standardized distribution which is determined with the z score. In the standardized distribution, the mean, median and mode are zero and you start increasing and decreasing z scores (to be discussed later).

How do we know if the data is skewed positively or negatively?

- Let us consider the yearly income of a family as an example, most of the employees are about 500Jds this is the mode which is the peak, but other people are about 20000Jds so these readings have shifted the mean to be bigger than the median and mode and this is called positive skewness the tail is on your right-hand **Mean>Median**. You might get a question what is the biggest value and the answer would be the mean in this scenario.

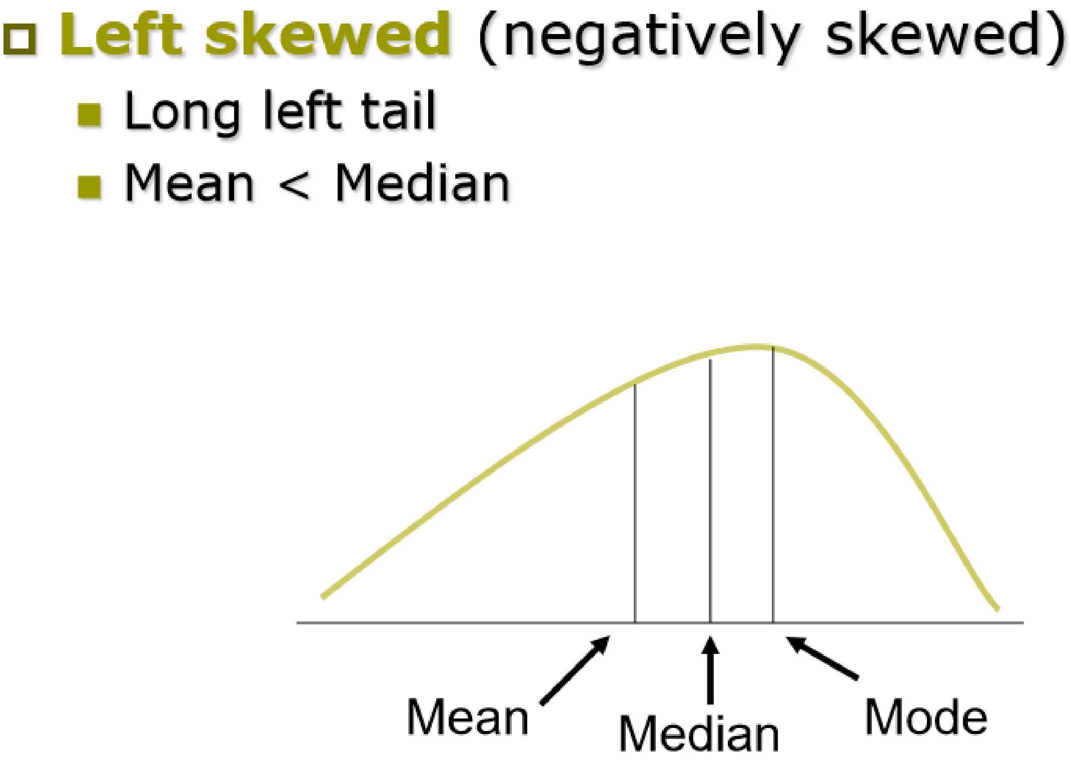


□ Right skewed (positively skewed)

- Long right tail
- Mean > Median

- If the curve is in opposite to the previous example; for example, the life expectancy of people in Jordan the mean is around 60 or

70 but the people who die at a young age pull the mean to the left. In this case the data is negatively skewed and the **mean<median**.

- 

□ Left skewed (negatively skewed)

- Long left tail
- Mean < Median

Measures of location consider the data at its midpoint in more than one method and of them are **measures of non-central tendency**.

Examples of commonly used quantiles:

- **Quartiles**.
- **Quintiles**.
- **Deciles**.
- **Percentiles**.

We will discuss **quartiles** and **percentiles** and the rest read them only for you.

For **Quartiles** consider this: you have a candy and there are 4 of you, how many cuts you should make? 3 cuts, to get 4 pieces.

The first cut is termed **Q1** and the second **Q2** and the third is **Q3**.

Percentiles to have a hundred part you should cut 99cuts.

in the exam, you are 120 and your result in the exam is 60 out of 100 you should be upset. but when you know that you are the third overall you will change and the marks are converted to percentiles you will calculate it by knowing the number behind you divide it by the total (in this example it will be (117/120) *100%=97.5% which is great). Usually software is used to calculate this. Excel or SPSS

Quartiles, how to know Q1, Q2, Q3?

Q1, **25th percentile**

Q2, the **median** and **50th percentile**

Q3, **75th percentile,** 75% of the observations are smaller than Q3 and 25% of the observations are larger.

How to calculate them? First, sort them from smallest to largest value.

If the values are **even** you divide by 4. (Ex: 12) you divide by 4 you'll have **3** if the result is an integer take the number after and divide by 2 Q1=((value3+value4)/2). If the result is not an integer you take the value after the number (Ex: 14) divide by 4 to get 3.5-> the Q1 is the 4th value

If the values are odd (ex: 11) you take the number after it for example 2.5 → Q1= the 3rd value.

The same goes for Q3 but form the other end.

these numbers represent the order of the value after sorting. See **slide 28 for more information.

Exercise in slide 29,

Original data: 3, 10, 2, 5, 9, 8, 7, 12, 10, 0, 4, 6

First sort the data: 0, 2, 3, 4, 5, 6, 7, 8, 9, 10, 10, 12

We have 12 values divide by 4 you'll get **3** take the 3rd and 4th values 3 and 4 "in this example" respectively you'll have **3.5** which is **Q1**. Now take 9 and 10 (by counting 3 from the other end) which are the 9th and 10th values you'll get **9.5** which is **Q3**.

How to know if it is positively or negatively skewed?

- You take the **biggest value** and subtract **Q3** from it in this example it is **2.5** and take **Q1** and the **smallest value** their difference is **3.5 3.5** is bigger than **2.5** so the data is **negatively skewed** if it was the **opposite situation then the data is positively skewed** and if **equal we would have symmetry.**
- To sum up: you look at **4** things (**Q1, Q3, the largest and smallest values**) and **subtract Q3 from largest** and **smallest from Q1**.
- **Max-Q3=X Q1-Min=Y**
- **X>Y → positively skewed X<Y → negatively skewed**
- **X=Y symmetric**
- Another important thing is **Q3-Q1** which is the **interquartile range**. These are **50%** of the values and an important application is to look at the middle class to see where it stands.

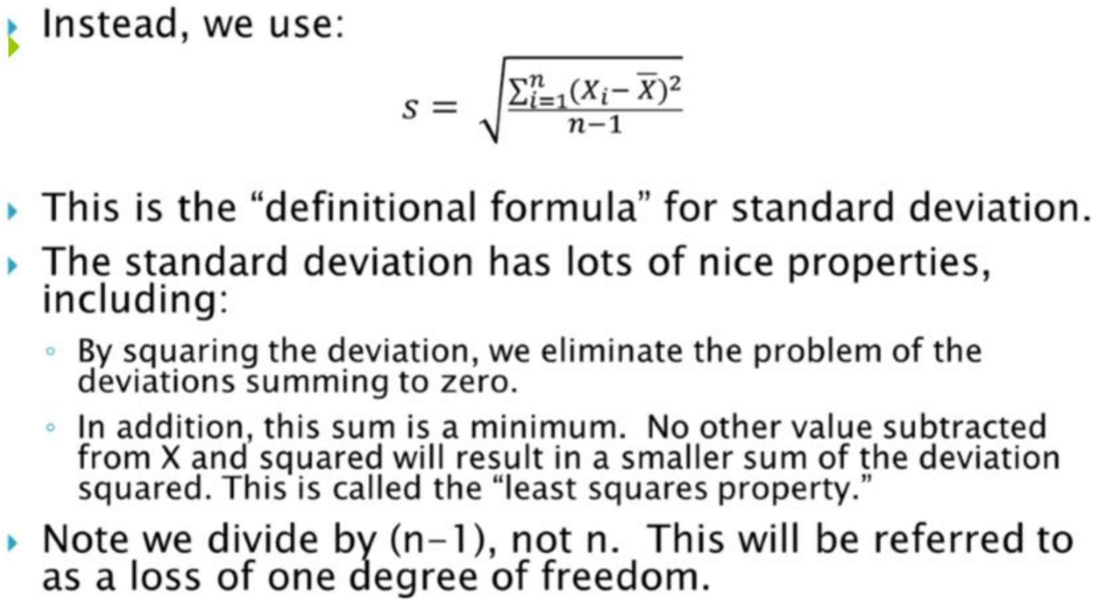
We will now consider how the data is variable, heterogeneous or homogenous and how are the values are away from the mean, how they are dispersed?

- First we will consider two values: standard deviation and variance and you must calculate the standard deviation to calculate the variance which is $(\text{St.d})^2$ we also need the range (highest-lowest) but it is affected by the extreme value so we need the **interquartile range "50% of the data" which is less effected by extremes**.
- If the things we are going to compare do not have the same units we must calculate **the coefficient of variation** which is the **(standard variation/mean).**

Standard deviation measures how are the values dispersed from the mean, the standard deviation for the **sample** is different from that of the **population**. To calculate:

$$\frac{\sum_{i=1}^{n}(X_i - \overline{X})}{n}$$

Average deviation, the average deviation value isn’t useful because it will sum up to 0.



- Instead, we use:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

- This is the “definitional formula” for standard deviation.
- The standard deviation has lots of nice properties, including:
  - By squaring the deviation, we eliminate the problem of the deviations summing to zero.
  - In addition, this sum is a minimum. No other value subtracted from X and squared will result in a smaller sum of the deviation squared. This is called the “least squares property.”
- Note we divide by (n−1), not n. This will be referred to as a loss of one degree of freedom.

We have the variety of choosing the last one, imagine you have 7 shirts every day to wear one the first day 7 options while the last day you only have 1 so we do not count it to have variety or a degree of freedom.

Example: st.d for the values 10, 20 ,30 ,50 ,70   Mean=36

$(10-36)^2 = 676$

$(20-36)^2 = 256$

$(30-36)^2 = 16$

$(50-36)^2 = 196$

$(70-36)^2 = 1156$

$$S = \sqrt{\frac{2300}{5-1}} = 23.98$$

See Slide 46 for huge importance!

Variance= S^2

Why this?

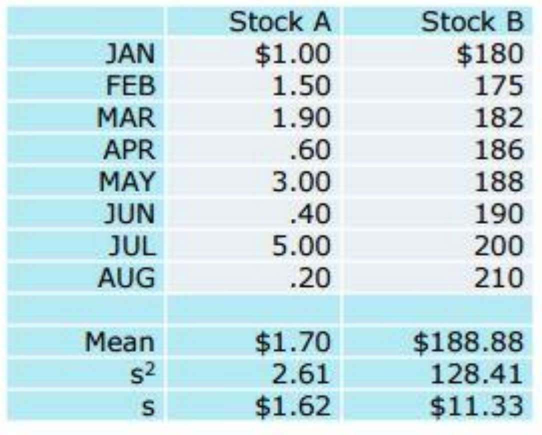
- The more heterogeneous the data the larger the S and the more homogenous the data the smaller the S the more normally distributed the data.

If the skewness index>1 you can’t work on parametric techniques. You will have to transfer the data meaning to use the geometric mean but if there is skewness you will have to change the data into ordinal data and use non-parametric techniques. **But there is limitation on generalization.**

Coefficient of variation: (st.d/mean)

$$CV = \frac{s}{\overline{X}}(100\%)$$

If we have 2 groups with different units we can’t compare them through S so we need to use CV. The closer it is to the 100 the more volatile the worst and it is even worse if more than 100 and the lesser it is the more homogenous the data. Consider this **slide 51**: two companies A and B deal with stocks and the price of a single share of A was 1dollar and that of B was 180dollars and in august it was .2 and 210 respectively. Which one is more stable?



	Stock A	Stock B
JAN	$1.00	$180
FEB	1.50	175
MAR	1.90	182
APR	.60	186
MAY	3.00	188
JUN	.40	190
JUL	5.00	200
AUG	.20	210
Mean	$1.70	$188.88
s^2	2.61	128.41
s	$1.62	$11.33

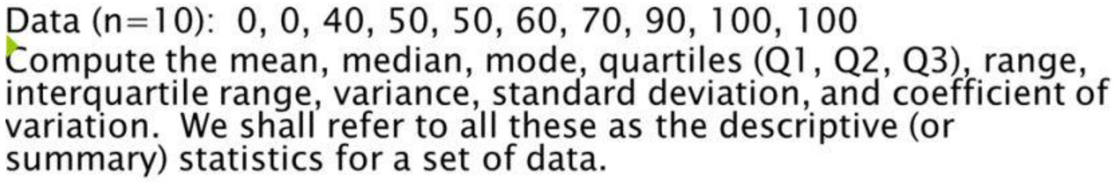
“We can’t compare according to S because the values are very different”

CV for A= 95.3%

CV for B= 6%

A is more volatile so it is not good (more heterogenous).

Now test yourself with this “exercise slide52 answer is there”:



Data (n=10): 0, 0, 40, 50, 50, 60, 70, 90, 100, 100

Compute the mean, median, mode, quartiles (Q1, Q2, Q3), range, interquartile range, variance, standard deviation, and coefficient of variation. We shall refer to all these as the descriptive (or summary) statistics for a set of data.

Q1-> 10/4=2.5 so Q1 is the 3rd value which is 40

Q3-> 3rd form the other side which is 90

Max-Q3=10

Q1-Min= 40

(Max-Q3) < (Q1-min) → negatively skewed

Interquartile range= 90-40=50

Range= 100-0=100

Mean= 56 modes= 0, 50, 100 median=((50+60)/2)=55

St.dv→ the sum of the square difference between the values and the mean

CV=st.d/mean