3

Biostatistics

Doctor 2018 | Medicine | JU

◉Sheet ○Slides

Done By

Raged Gharbia

Contributed in the Scientific Correction

Mohammad Abo Shaban

Contributed in the Grammatical Correction

Mohammad Abo Shaban

Doctor

Hamza Al-Duraidi

After providing you with the nice introduction of biostatistics, we are now ready to learn about the first step of biostatistics which is Descriptive Statistics.

We said before that the purpose of biostatistics is to organize and summarize the characteristics of the sample and to describe them with the minimal number of words and numbers.

Descriptive statistics have two main types:

a- Measure of location: ex the central tendency, (tendency toward certain location).

b- Measure of Dispersion: (tendency away from the central, away from each other toward limits).

In some resources they consider a third type which is the measure of shape: what does the distribution look like; symmetric or skewed to right or left.

Measures of location are two types:

a- CENTRAL TENDENCY: mean, median, mode

b- NON CENTRAL TENDENCY: quantiles which include quartiles, quintiles, percentiles, deciles.

NOTE: to make it clear when we are talking about a sample or population we use different letter, because of interpretation. For example, when you are talking about a sample mean you are sure 100% that the number is accurate (you calculate it), but in case with population it's an estimated number.

Whenever you see symbols in English letter → it's about a sample.
Whenever you see symbols in Latin letter → it's about population.

	Mean	Standard Deviation	Variance
Population	μ	σ	σ^2
Sample	$\bar{x}$	s	s^2

Measures of central tendency:

1) MEAN (average): the sample mean is the SUM of all the observations DIVIDED by the number of observation. The only problem with mean that it is easily influenced by extreme values (outliers) ☹ , imagine that you have the numbers 20,19,18,21 and 1, the last number (1) will decrease the mean and shift your distribution.

Generally, it's not a part of the data set. It's unique to the sample.

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \text{where } \Sigma X_i = X_1 + X_2 + X_3 + X_4 + \ldots + X_n$$

CALCULATE THE MEAN FOR (1,2,2,4,5,10)

THE MEAN = (1+2+2+4+5+10)/6=24/6=4

2) MEDIAN: it's the middle value of the ordered data, it is less influenced by extreme values. Also it's not a thing that you can calculate (like the mean), you find the median by FIRSTLY rearranging the data into an ordered array (ascending or descending, generally we order the data from lowest value to the highest value) then you find the number in the middle.(may be part of the data set)

Notice that half of the data are larger and half are smaller than the median .

If n is ODD, the median is the middle observation of the ordered array. If n is EVEN it is midway between the two central observations.

EXAMPLE

Note: Data has been ordered from lowest to highest. Since n is odd (n=7), the median is the (n+1)/2 ordered observation, or the 4^{th} observation. The median is 5, the mean is 32 (notice that the mean is influenced by the two extreme values 0, 100 while the median is not). The median is 5.

Q: What happens to the median if we change the 100 to 5,000?

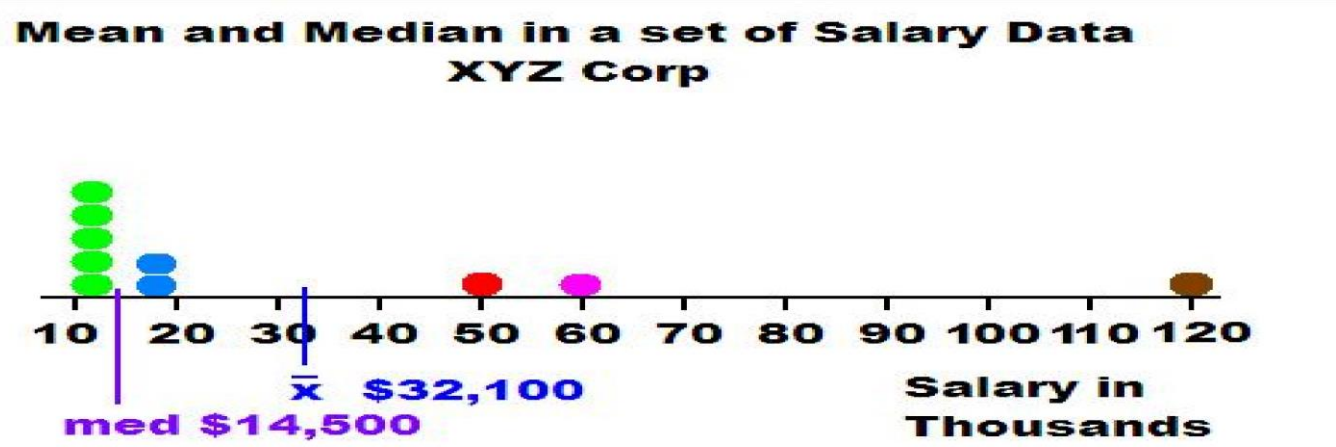
Not a thing, the median will still 5. Five is still the middle value of the data set. ☺

EXAMPLE

Note: Data has been ordered from lowest to highest. Since n is even (n=6), the median is the (n+1)/2 ordered observation, or the 3.5^{th} observation, *i.e.,* the average of observation 3 and observation 4. The median is 35, the mean is 35. (Notice that the mean=median, that indicate that the mean is a true mean; it is exactly in the middle).

HEY DIDDLE DIDDLE, THE MEDIAN IS MIDDLE

JUST ANOTHER EXAMPLE:

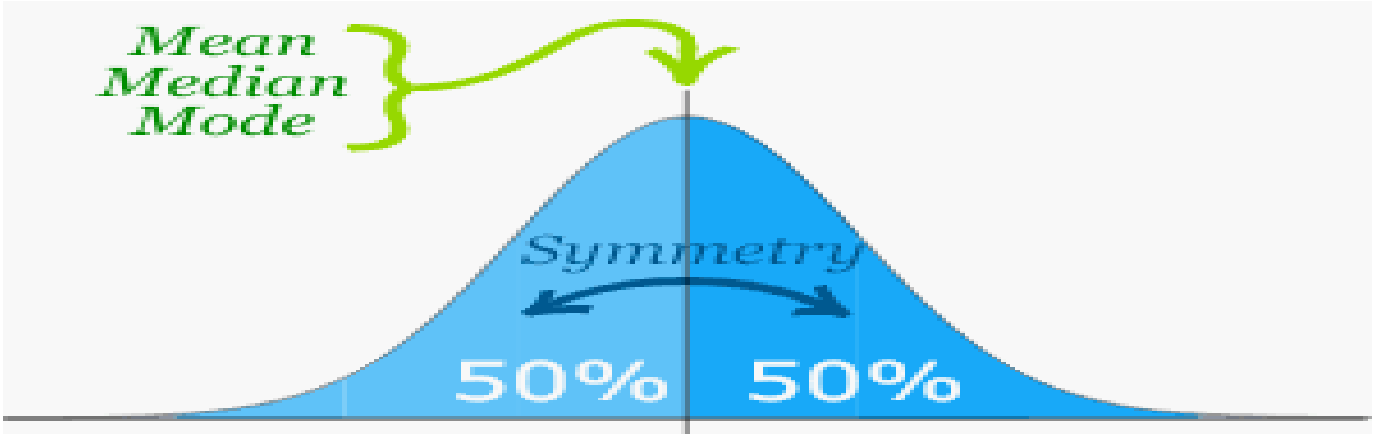


Every dot in the picture indicates a person's salary in a year, the first 5 salaries are 10 thousands, the 6^{th} and 7^{th} are almost 20 thousands, the 10^{th} person earns 120 thousands in a year. Notice that the last one will affect the mean and the distribution of the data. While the median is between the 5^{th} and 6^{th} person = (10+20)/2= 15 thousands. **But wait , the Mean is almost 32 thousands, which means that it's extremely influenced by the last person.**

3-MODE: it's the value of the data that occurs with the greatest frequency, and it is always part of data set , The mode in the previous picture is 10000. It's possible to have more than one mode (may not be unique), two mode we call it bimodal, three mode we call it trimodal. and we can observe no mode ex: 2,2,2,3,3,3,4,4,4 (all are observed equally).

Example. 1, 1, 1, 2, 3, 4, 5
Answer. The mode is 1 since it occurs three times. The other values each appear only once in the data set.
Example. 5, 5, 5, 6, 8, 10, 10, 10.
Answer. The mode is: 5, 10.
There are two modes. This is a ***bi-modal*** dataset.

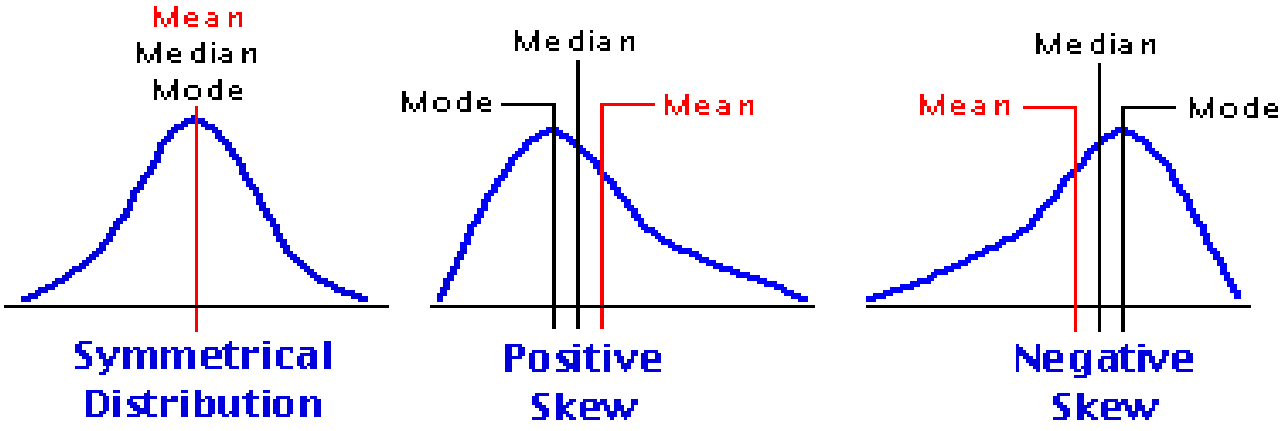
We have discussed the three measures of central tendency: Mean, Median and Mode. But what if these three measures are equal in a certain sample … then the sample will be distributed in a symmetric manner, in other words, it will be distributed normally. (**Notice that Right and Left sides are mirror images**)



But what if these three measures are not equal? Well, we have two situations:

1- Right skewed (positively skewed):
Here the Mean>Median>Mode.
This distribution has long right tale and is influenced with extreme high values.

2- Left skewed (negatively skewed):
Here the Mode>Median>Mean.
This distribution has long left tale and is influenced with extreme low values.



Notice that the median in the skewed distribution is always in the middle between mean and mode.

We have discussed the measures of central location, now let's shift to the measures of non-central location (**quantiles**):

1- Quartiles: الأرباع Q

We split the ***ordered*** data into **four** equal parts, in order to split it we make THREE cuts (imagine it ☺).

- The site of the first cut is called Q1 or the first quartile:

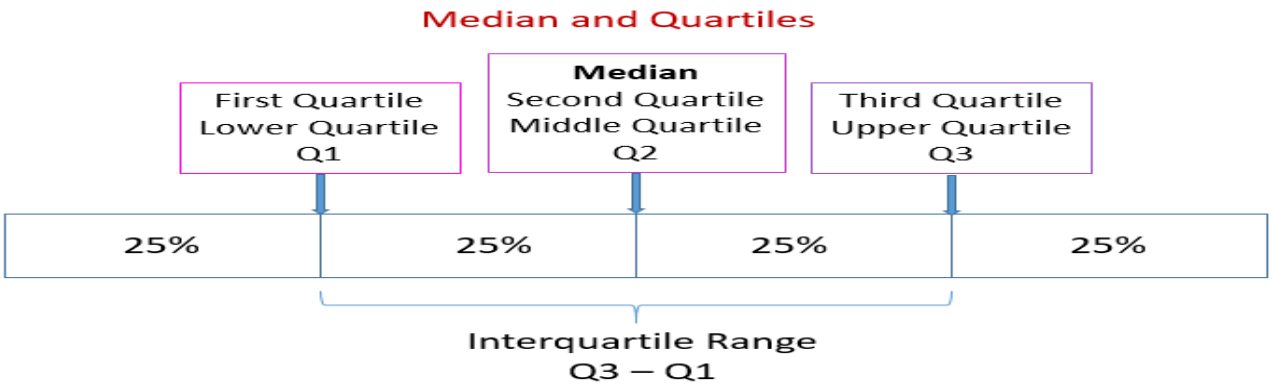
*25% of the data are smaller than Q1, 75% are larger.

-The site of the second cut is called Q2 or the second quartile:

*50% of the data are smaller than Q2, 50% are larger, which is the median ☺. (Median = Q2)

-The site of the third cut is called Q3 or the third quartile:

*75% of the data are smaller than Q3, 25% are larger.



2- Quintiles: الأخماس Qn/QN/QU

We split the ***ordered*** data into **five** equal parts, in order to split it we make FOUR cuts:

-The site of the first cut is called Qn1:

*20% of the data are smaller than Qn1.

-The site of the second cut is called Qn2:

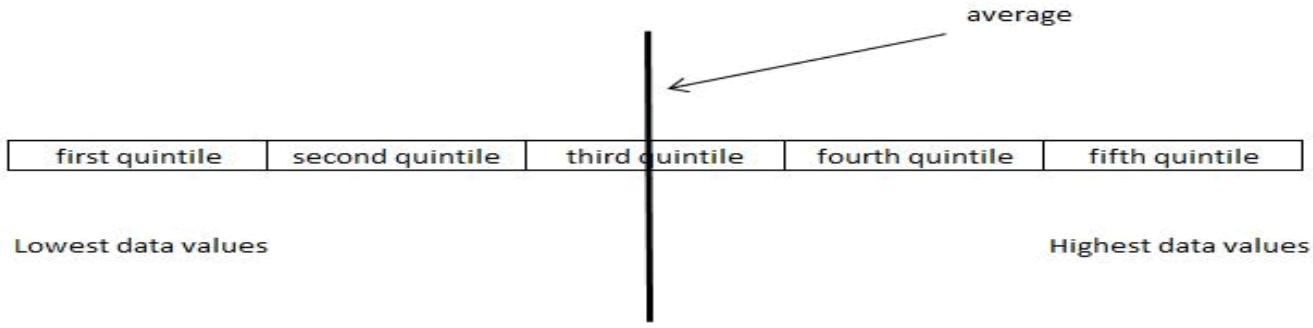
*40% of the data are smaller than Qn2.

-The site of the third cut is called Qn3:

*60% of the data are smaller than Qn3.

-The site of the fourth cut is called Qn4:

*80% of the data are smaller than Qn4.



3-Deciles: الأعشار D

We split the ***ordered*** data into **ten** equal parts, in order to split it we have to make NINE cuts.

*The first cut is called D1, 10% of the data are smaller than D1, so on with the rest Deciles.

4-Percentiles: النسب P

We split the ***ordered*** data into **hundred** equal parts, in order to split it we have to make 99 cuts.

*This measure is recommended when you have a LARGE sample.

-The site of the first cut is called P1, 1% of the data is smaller than it, and so on. (I won't mention all percentiles, it will take ages xD)

SOME IMPORTANT NOTES:

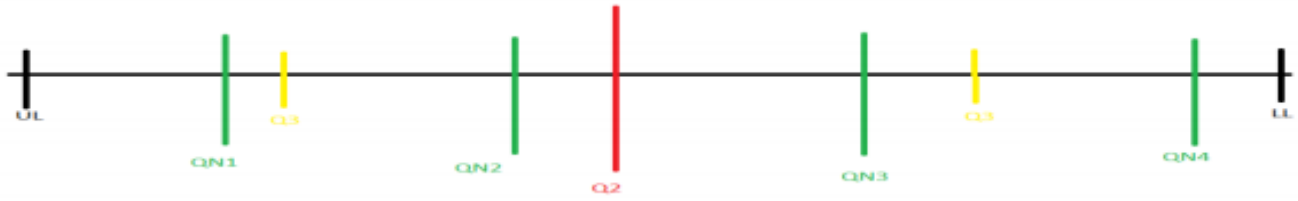
1- We don't have Q4, Qn5, D10 and P100, because it's an Upper Limit.

2- The distance between the first Quartile (Q1) and the third Quartile (Q3) is called IQR which is one of the measures of Dispersion. (will be discussed later).

3- Notice that:

Median = Q2 = D5 = P50 / Qn1 = D2 / P20 = D2 = Qn1.

4- For each sample we have Upper limit (the highest value) and Lower Limit (the lowest value), The difference between them is called the Range, which is also a measure of dispersion.



5- There is an easy way to find Q1 and Q3 of the sample by:

A- Order the sample and find their median (remember that the median divide the sample into two equal parts)

B- Take the first part of the sample which contains the numbers smaller than the median and find their median, that is Q1.

C- Do the same thing for the second part and you will find Q3.

THERE ARE OTHER INFORMATION AND EXAMPLES IN THE SLIDES THAT WERENT DISCUSSED BY THE DOCTOR, PLEASE CHECK THEM.