Biostatistics

Doctor 2017 | Medicine | JU

Number >>	5
Doctor	Hamza Aduraidi
Done By	Rawan Almujaibal
CorrectedBy	Dana alqatawneh

Sheet has been written from section 3 record

In the previous lecture, we have talked about the process of hypothesis testing. We discussed briefly what biostatistics is, and we defined it as a branch of applied math that deals with collecting, organizing, and interpreting data using well-defined procedures. There are 2 types of biostatistics:

Descriptive statistics: it involves organizing, summarizing & displaying data to make them more understandable.

Inferential statistics: it reports the degree of confidence of the sample statistic that predicts the value of the population parameter.

Some other definitions:

Data: any type of information.

Raw data: data collected as received. numbers not summarized and organized.

Organize data: data organized either in ascending, descending or in a grouped data.

Descriptive Biostatistics

The purpose of descriptive biostatistics is to make our huge sums of data in our sample more meaningful and manageable. For example, if a study involves 100 students asking them about how many meals they have a day. The result of the study is collected and organized to be like this "the mean of their answers was 3 meals a day". Instead of saying student no.1 has 4 meals; student no.2 has 3 meals and so on. Hence, it is summarizing the data.

SLIDE:

Descriptive Measures

- *A descriptive measure is a single number that is used to describe a set of data.*
- *Descriptive measures include measures of central tendency and measures of dispersion.*

Descriptive biostatistics has three types:

1- **Measures of location** (معايير أو مقاييس التموضع):
   It measures the tendency of the sample towards a certain location.
   2 types:
   -Measures of central tendency, such as Mean; Median; Mode (towards the centre).
   -measures of non-central tendency, such as quintiles (the location of the sample either less than 25% or more than 75%).
2- **Measures of dispersion** (مقاييس التشتت):
   It measures the dispersion individuals in the sample, and how deviated they are from each other.
   For example, Range, Interquartile range, Variance, Standard deviation, and coefficient of variation.

3- **Measures of shape** (the least important measure):
When we draw the individual of a sample on X axis and Y axis we get a draw called "distribution". The distribution could be symmetrical (موزعه بشكل متساوي) or skewed(بإتجاه واحد).

1- Measures of location

It is the most important measure especially of the measures of central tendency and we use it often in our life.

1- **The mean (average) is:**

The sample mean is the sum of all the observations (ΣX_i) divided by the number of observations (n):

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

NOTE: the symbols of the population we use Latin symbols, and symbols of the sample we use English letters.

Back to our example above about the study which asks the students how many meals they have a day. To make the results very easy and meaningful we calculate the mean value. If we want to calculate the mean value of the sample: the sum of all the observations (collect the answers of the students) divided by the number of the observation (100 students or the participants).
N=sample size (100 in this example)

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

μ = population mean
Σ = summation sign
x_i = value of element i of the sample
N = population size

NOTE: there is another name for the mean called ***sample average*** or ***arithmetic mean***.

There is only one problem we should put it in our consideration when we calculate the mean; it is very sensitive and easily affected by extremely large or small values at the very end of the spectrum.

Applying this to our example:
Let's say most of the students answered that they have between 2 to 4 meals per day and one of the 100 students said he has 12 meals per day. All these results will be added to the spectrum including the one who is having 12 meals per day as well. If we apply the equation above will get very big value that is influenced by the extreme value (12 meals per day).
So with the extreme values, the mean would not be a really good expression or representation of the whole sample.
What to do in this case?

We are going to calculate the median instead of the mean.

2- Median:

It is not calculated, but found. It is the middle value of the *ordered* data. To get the median, we must first rearrange the data into an ordered array (in ascending or descending order). Generally, we order the data from the lowest value to the highest value. Therefore, the median is the data value such that half of the observations are larger and half are smaller. It is also the 50^{th} percentile. If (n) is odd, the median is the middle observation of the ordered array. If the (n) is even, it is midway between the two central observations.

Back to our study; we will see the 100 answers and we look for the middle value (or located in the centre) of these answers which is the value 50 and 51 after we order them. We get the value of participant number 50 and add it to the value of 51 participants, and then we divide the resultant value by 2. This would be our median.

The median has 3 interesting characteristics:

1- The median is not affected by extreme values, but only by the number of observations.
2- Any observation selected at random is just as likely to be greater than the median as less than median.
3- Summation of the absolute value of the differences about the median is a minimum (from the slide)

Why is the median important in some scenarios instead of mean?

Because it's free from the huge influence of the outliers, so it gives us better estimation of the central tendency of that sample.

-No matter how many times a single observation is repeated, it still is count as one, and when you calculate the median it respects each individual reading as its own value even if it's repeated.

-both the Median and Mean is important, none of them is more important than the other. They measure different things, and it's better to combine both to get a full view of how much tendency toward the center a certain sample has.

- if a sample contains observations that are either extremely high or extremely low, the median will either fall near one of them. This doesn't represent the data fairly. So, the best solution is to calculate both the mean and the median and view the sample more clearly.

Some mathematical examples:

First example: (0, 2, 3, 5, 20, 99, 100).

We have 7 people; person no. 1=0, person no.2=2, person no.3=3. Person no.4=5, etc…

First thing we do is ordering the data then we find the middle value (we have to know the numbers are they even or odd)

Note: data has been ordered from the lowest to highest already. Since (n) is odd (n=7) the median in the (n+1)/2 ordered observation or the 4^{th} observation.

Answer: the median is 5.

Note: what happens to the median if we change the 100 to 5000? Nothing the median will still be 5. Five is still the middle value of the data set.

If you calculate the mean, it will be nearly 32 (have you understand now what we mean by it is getting affected by extreme values).

Second example: (10, 20, 30, 40, 50, 60).

Note: Data has been ordered from the 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.

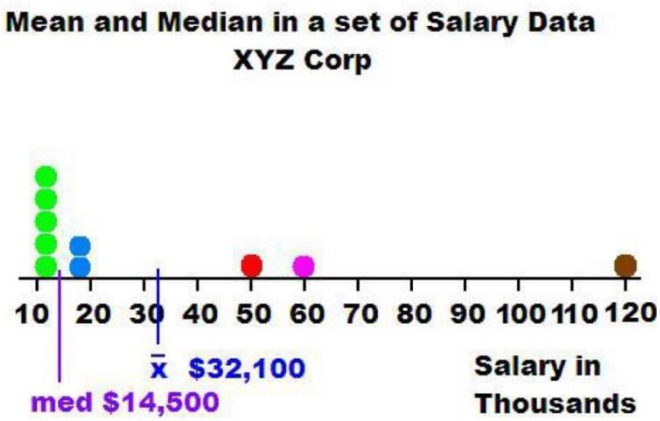
= (30+40)/2

Answer: the median is 35.

If you calculate the mean, it will be nearly 35 (so we can conclude from that we don't have extreme values neither lowest nor highest values; the mean will be very close or equal to the median values). The sample is equally distributed (symmetric distribution).

From the slide:

- *Advantages of the median:*
  The median is less affected by extreme values.
- *Disadvantages of the median:*
  The median takes no account of the precise magnitude of most of the observations and in therefore less efficient than the mean.
  If two groups of data are pooled the median of the combined group can't be expressed in terms of the medians of the two original groups but the sample mean can.



3-Mode:

It is not calculated, but found. The Mode is the single value that has been repeated the most. Unstable index: values of modes tend to fluctuate from one sample to another drawn from the same population.

Let's have this example:

Salaries were: ($10000, $10000, $10000, $10000, $10000, $20000, $20000, $50000, $60000, and $120000).

We have noticed that salary $10000 has been repeated 5 times, so
The answer: the mode is $10000.

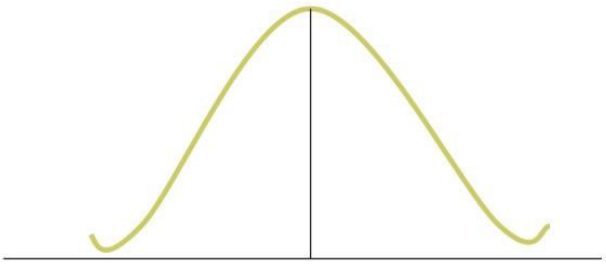
Note: some samples have one mode (uni-mode) as in our previous example; other examples may have more than one mode (2 modes called bi-modal) (more than 2 modes called multi modal), or even sometimes we don't have mode at all.
The mode is different from the mean and the median in that those measures always exist and are always unique. For any numeric data set there will be one mean and one median.

Another example: (5, 5, 5, 6, 8, 10, 10, 10).
Simply the answer: the mode is 5 and 10
There are 2 modes. This is bi-modal dataset.

Why is Mode important?
Because the most frequent value seems to be an important value, and I need to focus on it.

From the slide:
Comparison of the mode, the median, and the mean:

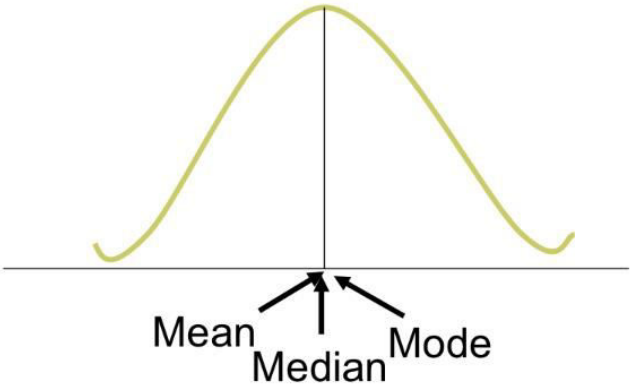
- *In a normal distribution, the mode, the median, and the mean have the same value.*
- *The mean is the widely reported index of central tendency for variables measured on an interval and ratio scale.*
- *The mean takes each and every score into account.*
- *It also the most stable index of central tendency and thus yields the most reliable estimate of the central tendency of the population.*
- *The mean is always pulled in the direction of the long tail, that is, in the direction of the extreme scores.*
- *For the variables that positively skewed (like income), the mean is higher than the mode or the median. For negatively skewed variables (like age at death) the mean is lower.*
- *When there are extreme values in the distribution (even if it is approximately normal), researchers sometimes report means that have been adjusted for outliers.*
- *To adjust means one must discard a fixed percentage (5%) of the extreme values from either end of the distribution.*



Now we've finished the measures of central tendency and after we move on to the next concept let have this **example;**

If we have a sample and its mean = 10, median= 10, and the mode = 10. What do you understand from these values?

As it is obvious that the most reading were in number 10, because mode =10. And we understand as well that the distribution on the right is equal the one on the left, because of the median= 10. Also, because the median value is equal the mean value that means there are not extreme values.
So we get this sort of distribution as it is shown in the diagram.



So, knowing the mean, median and the mode help us to figure out the shape of the distribution.

Distribution characteristics:

1- In our previous example is called **normally distributed**, because you find the observations in the middle have the highest frequency.

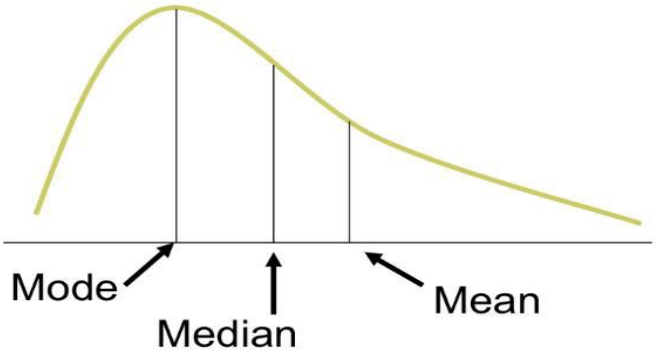
In the normally distributed graph:

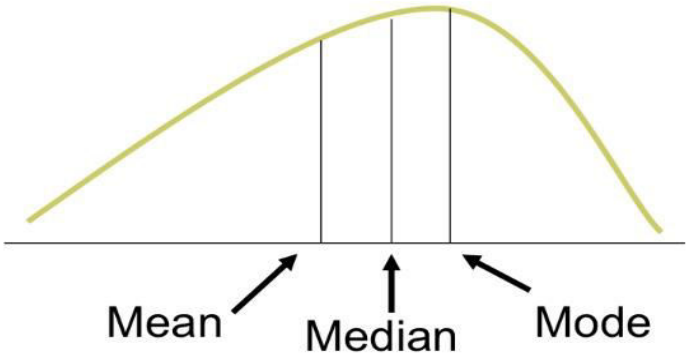
- Median is in the mid-point as the right side is exactly equal to the left side.
- If you sum the values and divide them on their number you'll find that the mean is in the mid-point.
- And the most frequent value (mode) is in the mid-point.

The conclusion: in graphs which are normally distributed (hill shaped); median is equal to both mean and mode (all are at the same point). it's great to see this graph in statistics because it's normally distributed at the sample level and most likely to be normally distributed at the population level too. So, when you draw inferences about the population they are most likely to be accurate.
This is called "symmetry"; because the value in the middle divide the curve into two identical sides, and this value is equal to median, mode and mean.

2- Let say if mean= 15, median 10, and mode =12 how does that looks like in shape?
In this curve, people seem to have values more to one side than the other (high frequent observations concentrated which the mode is at left side).
We get not symmetrical distribution (skewness); we called the distribution in this case right skewed.

This type of curve either is right skewed (positively skewed) or left skewed (negatively skewed), BUT WHAT DETERMINES IF IT'S RIGHT OR LEFT SKEWED?
The direction of the tail; if the tail is directed to the right the curve will exhibit "right skewed distribution". If the tail is directed to the left the curve will exhibit "left skewed distribution".
In both scenarios (right and left skewed distribution) we broke the symmetry rule, so mean, median and mode are not equal, and they have different values; mode value is the peak (most frequent). Median values are in the middle and mean depends on the summation of the values divided by their number (balance point).





Note: The mean is always pulled in the direction of the long tail, that is, in the direction of the extreme scores.

Left skewed	Right skewed
Negatively skewed	Positively skewed
The mean is lower than the mode or the median	The mean is higher than the mode or the median

Best of luck….