Biostatistics

Doctor 2017 | Medicine | JU

Number >>	7
Doctor	Hamza Alduraidi
Done By	Leen Alsahele
CorrectedBy	Farah Maayah

In the previous lecture, we finished talking about all **measures of location,** whether measures of **central tendency** (mainly: Mean, Median and Mode). Or, measures of other location of distribution of variant (**non-central tendency**) such as: Quartiles, Quintiles, Deciles, Percentiles.

There is something common between all those ***measures of location*** that is all measure the sample members' tendency toward a certain location (center, or elsewhere). This is the ***first type*** of descriptive statistics.

The second type of descriptive statistics is called **measures of dispersion**

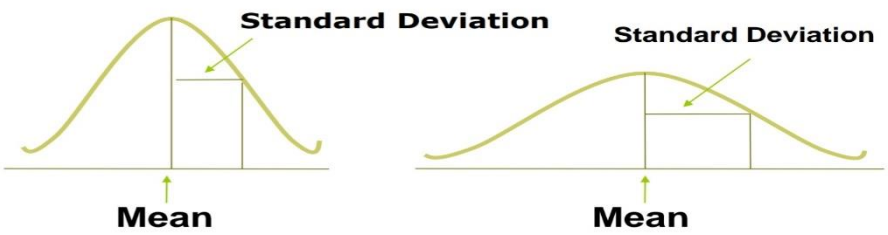
While all measures of locations try to assess the sample members' tendency towards each other (or towards certain location on the line), these new measures of dispersion are the **opposite**. They try to measure sample members' **tendency towards the outside** (away from each other-away from the center-towards the right and the left- how spread out the scores are). It is either homogeneous or heterogeneous sample (from slides).

measures of dispersion are the measures that try to asses to what extend does each member of the sample "each value of X " tend to be located away from the center or away from the other values of X

These measures of dispersion have many examples:

1- Range
2- Standard deviation
3- Variance (the square of standard deviation)
4- Coefficient of Variation (CV)
5- Interquartile Range
6- Relative Standing

Dispersion as a concept is represented in these two graphs.



These two graphs represented the distribution of 2 different samples, each of these 2 samples is symmetric, and each of them has a mean and median located at the same point. But, one of them is **narrower** than the other or one of these

distributions is **wider** than the other. The notion of the width of distribution means how much dispersion there is in that sample.

The wider the distribution of certain sample is, the larger the dispersion is.
Meaning that the mean distance between the sample members from each other and the mean distance between each sample member and the mean can be narrow (as the left picture) or wide (as the right picture).

This is the concept that we are trying to assess in measures of dispersion (to what extends do the members of the sample tend to be located away from each other and away from the sample. To measure this, we have several types of measures the first and the simplest one is called **The Range.**

1-The Range

It is basically the distance between the **upper limit** (highest value of x, located to very far right) and the **lower limit** (lowest value of x, located all the way to the left).

Range= X (upper limit) – X (lower limit)

In the last picture, the right sample has wider range
Unfortunately, similar to the mean, the range is very sensitive to extreme values, either to the right or to the left or both.
If we have a sample of 100 students, and all students got a grade between 5 & 7/10, but 1 student is "قطاعة " got 10/10. This student widens the range to the right. Other student "بايعها" got 0 → extreme value to the left. These 2 students increase the range.

To solve this problem "sensitivity of the range", there is other type of measures of dispersion that is called The Standard Deviation.

2-The Standard deviation

It is the most important, meaningful and useful type of measures of dispersion. In fact, it is one that is reported more than any other measures of dispersion in researches.

So, why is standard deviation a genius number to report?
Because standard deviation is something that we calculate using an equation

This is the calculation equation for standard deviation.

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

We use the letter **S** to represent standard deviation if we talk about a **sample**.
But if we talk about a **population**, we use **sigma** instead.
So, this equation is what we use to calculate the standard deviation for any sample we have, since we have all values of x and we know the sample size, we can calculate it.
As you see, x is part of the equation. So, to calculate S, you should calculate the mean before.

What does the standard deviation tell me exactly?

It tells me what is the average distance between each of the values of x and the mean (it measures the average distance of each individual sample members from the mean)
If u have a sample with the mean 3, and you calculate the standard deviation using the equation above →S=1. What does that mean?
It means that each value of x (each member of the sample) is located in average one unit away from the mean, either to the right or to the left.
This indicates how much individual sample members are far away from the center.

In this example, we have 2 samples (x&y). Each has 5 members and both have the same mean (=3).
For X = 1+2+3+4+5/5 = 3
For Y = 0+0+0+5+10/5 = 3

X_i	Y_i
1	0
2	0
3	0
4	5
5	10

You may guess that they look similar because the mean is similar; when in fact, they look nothing similar.

To make it clearer, let us calculate the standard deviation of each. When applying the equation, the results were:

Standard deviation of x=1.58, and of y=4.47.

Sample y has deviation/dispersion/variability 3 times more than sample x.

What do these 2 numbers mean?

In sample x -> any given individual is located in average 1.5 units away from the mean.

In sample Y -> any given individual is located 4.5 units away from the mean (greater variability of the values of y).

X	$\bar{X}$	$(X-\bar{X})$	$(X-\bar{X})^2$
1	3	-2	4
2	3	-1	1
3	3	0	0
4	3	1	1
5	3	2	4
		$\Sigma=0$	10

Y	$\bar{Y}$	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
0	3	-3	9
0	3	-3	9
0	3	-3	9
5	3	2	4
10	3	7	49
		$\Sigma=0$	80

[Check these results with your calculator.]

The question is, do we in research and biostatistics prefer (desire) a greater or a narrower standard deviation?

A narrower. Why? Because when you have a smaller standard deviation, you have less variability and dispersion. Then, the mean is a good representation of everybody in the sample. But when the standard deviation is high, the mean is not a good enough representation of everybody in the sample.

Let's take another example => let's say that you are a YouTuber and you have 1,000 subscribers. And another YouTuber has another channel with 1,000 subscribers as well.

The mean age of the first channel subscribers is 30 years old and the mean age of the second channel subscribers is 30 years old as well.

The standard deviation of the first Sample is 1. But the standard deviation of the second sample is 10.

That means that the ages of subscribers of the first channel are more condensed (close to the mean & near to each other) => the distance between each one of the subscribers and the mean "30 years old" is in average 1 year old. The age of people in this sample -> as old as 31 or a little older & as young as 29 or a little bit younger.

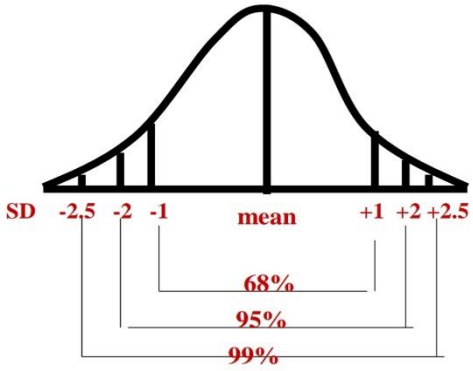
But in the second sample, there is more divergent in terms of the age (as old as 40 or even older & as young as 20 or even younger).

The average distance between any given person in terms of the age and the mean on the first sample is one year, in the second is as high as 10 years.

So the YouTuber that has the first channel can decide what type of content he need to post. However, the second one has higher variety in terms of age, Therefore, the decision in terms of content will be harder.

So, we always desire a homogeneous distribution, where people are similar enough to the mean. **That's why we always prefer a small standard deviation rather than a large one.**

The statistician around the world, by using probabilities, they got these percentages in any sample that is **large enough** and **normally "symmetrically" distributed**:



- If I move 1 SD to the left of the mean & 1 standard deviation to the right of the mean, this range contains **68%** of all sample members. (In the 1st sample in the previous example "YOUTUBE CHANNEL", 68% of all subscribers are aged between 29&31~~~ in the 2nd sample, 68% are between 20&40).

- If I move 2 SD to the right & 2 SD to the left, I would have covered **95%** of the members (95% of subscribers are between 28 & 32 years old ~~~ in the 2nd, 95% between 10&50).

If I move 2.5 SD to the right & 2.5 SD to the left, **99%** of individuals are covered (between 27.5&32.5 in the 1st sample~~~between 5&55 in the 2nd).

Only **1%** of individuals of any normally distributed sample are located beyond 2.5 SD to either way. In the first channel, you may find a 45 years old subscriber or 11 years old child, but these are a minority and may not be more than 1% of the sample.

If we know that these percentages are true for every sample in the world as long as it is normally and symmetrically distributed, this will be very useful in predicting people's health and health status, like cholesterol, nicotine, sugar levels in the blood and many other health related characteristics.

3- Variance

Another measure of dispersion, which is simply the square of standard deviation **(s^2 in a sample and σ^2 in a population).**

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

All I need to know that is just the square of SD and it's one of the examples of measures of dispersion. But how do I use it? We will know when we talk about inferential statistics and we will realize that variance is important to calculate something called standard error (part of calculations related to T test, Chi square and other)

4- Coefficient of Variation (CV)

The equation is:

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

The higher the CV is, the higher the dispersion is. The lower the CV is, the lower the dispersion is.

How can the doctor ask about it on the exam? Simply, giving us numbers and ask to calculate the mean, the SD and then the CV.

5- Interquartile Range (IQR)

IQR = value of x on Q3 – the value of x on Q1

The range is always greater than IQR because range covers 100% of all values of x, while IQR covers only the middle 50% because it leaves out the last quarter to the right and the last quarter to the left.

How can the doctor ask about it on the exam? Also, giving us values of x and ask to find the median, Q1, Q3, probably some percentiles, some deciles, then ask to find IQR.

Example: A class had an exam of 30 marks. The highest mark was 29, the lowest mark was 9. The person that was located on Q3 got 21, and the person that was located on Q1 got 12.

1) The range.

Range = 29-9 = 20

2) IQR= 21-12 =9

The five number summary

It is a nice way of summarizing any sample out there. It is done by providing only 5 numbers.

These numbers are:

1) Lower limit

2) Q1

3) Median (Q2)

4) Q3

5) Upper limit

If we have these five numbers, we will get a nice picture of what the distribution may look like.

Example1: I have the following five numbers summary: 2, 4, 6,8,10

What is the range? 10-2=8

What is the IQR? 8-4= 4

What is the median? 6

Do you think that this distribution is symmetrically distributed?

Yes it is, because the distances are converged and all values that are right to the median and all to the left cover the same distance.

Example2: 2,4,6,8,50 (one extreme value to the right)

Range= 48

IQR=4

Do you think that this is nicely symmetrically distributed sample?

No. What type of distribution is it? **Skewed to the right** (the extreme value makes a tail to the right).

Example3: 2, 18,20,22,24

What type of distribution is this? **Skewed to the left** (the extreme value dragged the distribution all the way to the left, makes a tail to the left).

هاد الشيت للي استناه كتير، أما الي راح وسمع الريكورد فلا شيت له 😊😊😊😊😊