◉Sheet ○Slides

Number

5

Done by:

Sarah Alda'jah

corrected by:

Omayma Hassanin

Doctor

Mahmoud hussami

Descriptive Statistics

Statistics is branched into descriptive and inferential statistics.

Descriptive statistics is all about summarizing the data to understand the whole picture that upon it you will build your study. (Summarizing is the best way to deal with data)

You start with raw data (raw data that has been collected and not dealt with), then you organize it and you clean it up and dispose of the contaminated data, outliers* & errors.

For example: in a questionnaire concerning the heights of the individuals in the sample, one of the answers given were 3m; this is not a real value and you have to clean it.

*Outliers [extreme values]: are values that are far away from the range and we don't compute such values in the mean because they would result in a false reading. (It pulls the mean off)

How do you clean your data?

By descriptive measures. *10-20min

Once you collect the data, it's called raw data. Then you organize it and input it into one of the statistics softwares such as excel and SPSS.

Data is any type of information.

Organized data is data organized either in ascending, descending, or in a grouped data.

(Numbers are not always considered data; they need to be processed and worked on, and sometimes it is enough just the way they are. It all depends on the type of study you are working on.)

Input number ⟶ Output information. You insert factors and you come up with results.

One model of dealing with health care issues is managemental model; the story here is different, the data you collect is considered information.

For example: the height of an individual is 170cm → this is info.

You have 5JD today → this also a piece of information.

*in the next paragraph the doctor talks briefly about excluding and including criteria *:

If you want to calculate the mean of the allowance of Jordanian 2nd year medical students per day, you have to put excluding criteria that disqualifies any one that's not Jordanian from being interviewed for this study in order to get the right information. And if some students receive a weekly or monthly allowance, then you have to divide it by 7,30 respectively.

Talking about excluding and including criteria furthermore, if you want to do a research that concerns the residing citizens of Jordan, then you better not do it in the summer because it would include the tourists and immigrants whom would have a bad impact on your study.

So, when doing a study, you have to look out for the extraneous variables and the confounding variables.

Descriptive Measures:

A descriptive measure is a single number that is used to describe a set of data. And it includes three types:

1- Measures of central location
2- Measures of variability
3- Measures of shape: by this measure we can know if the data is normally distributed, if there is any skewness, and how much the index of skewness is.

A) Measures of central location:

They look at the data upon the central point; it is a property of the data that they tend to be clustered about a center point.

And they are two types:

- **Central tendency measures**
- **Non-central tendency measure**

Central tendency measures:

Measures of central tendency (i.e., central location) help find the approximate center of the dataset.

Central tendency measures are called 'the average' in slang language, but in biostatistics there's nothing really called the average. Instead, there are median, mode and mean,

and they are different from each other, so we can't call these measures a one name that is average especially when you are dealing with skewed data.

So, the central tendency measures include:

1- The mean (generally not part of the data set)
2- The mode (always part of the data set)
3- The median (may be part of the data set)

The mean and median don't always have to be within the data set.

For example: you were given these numbers and were asked to calculate the mean!

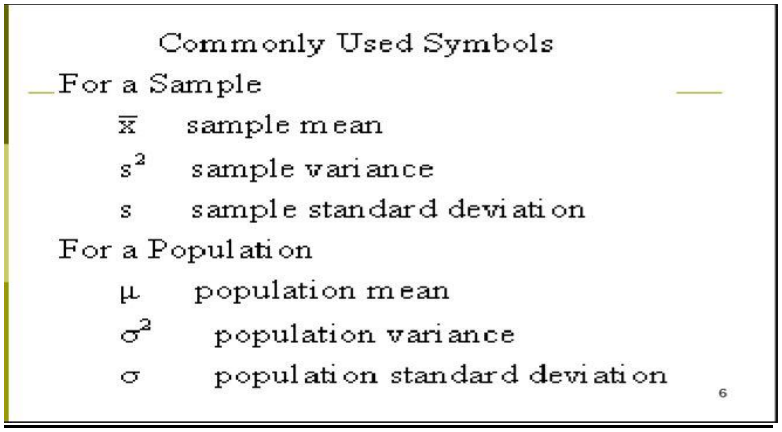
1,4,6,3. the mean is (1+4+6+3)/4=3.25 which is not within the set of data given.

Now for the median, if the given values were even, then it's not within the data set and if they were odd, then it would be within it.

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

There's only one case where they would all be the same value. That is when the data is normally distributed and the skewness is zero. (Symmetric data)

Commonly Used Symbols:



Non-Central tendency measures:

Here we deal with a cut in the data line.

In this case, we call them quantiles; which are cut points in the data line. If you need percentiles or if you require charts then we use this kind of measure.

Back to central tendency measures.

a) The mean: (arithmetic and geometric mean)

The mean is the arithmetic average of the distribution and the measure of central tendency with which most people are familiar. The mean is easy to calculate [Simplicity]. The mean is most appropriately used to describe rational interval-level data, but in some cases, it may also be used to describe ordinal data. To calculate the mean, you add up the values of each observation in the data set and then divide by the number of observations.

Sample Mean	Population Mean
$\bar{x} = \frac{\Sigma x}{n}$	$\mu = \frac{\Sigma x}{N}$

where Σx is sum of all data values

N is number of data items in population

n is number of data items in sample

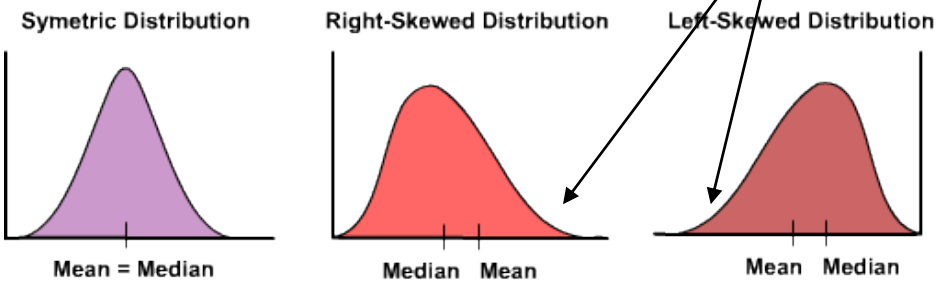
In the formula to find the mean, we use the "summation sign" — $\sum$ ν
This is just mathematical shorthand for "add up all of the observations".
*20-30 min
Mean for the sample = X or M, Mean for population = mew (μ).

The mean is very sensitive to the extreme values because these values either pull the mean negatively or positively, and this could affect decision making.
So, you should trim the data, you should take the first and last 5% of the mean calculation, so we're left with the mean of 90% of the population [Trimmed mean].

For every set of data there's only one mean so it's considered unique. The mean could be within the set or out of the set of data.

*Note: In order for us to say that the data is normally distributed, the standard deviation above and below the mean should be near 68%. If the previous requirement was not achieved, data should be transferred first; in this case, we use geometric mean instead of arithmetic mean.

Skewed data is recognized by the presence of a tail either on the right or left side.



**If the data is skewed we use the median

b) The median:

The median is the value that is in the middle of the distribution; it is also called the 50th percentile. It is the data value such that half of the observations are larger and half are smaller.

One way to find the median value is to 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 and find the value that is in the exact middle of the distribution.

There are two cases of median calculation; if the values are odd and if they are even.

*Even: If the median of a distribution with an even number of values must be computed, the two values in the middle of the distribution are averaged (mean).

*Odd: there will be only one middle value once they are ordered so no need for averaging here.

The median is not affected by extreme values and it's used in the case of both ordinal and skewed data.

The mean and the median are unique for a given set of data. There will be exactly one mean and one median.

*Note: if the data is continuous and skewed, we use the non-parametric measures and we convert it to ordinal level.

- The 3 characteristics of the median :
  1- The median is not affected by extreme values, 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 the median.
  3- Summation of the absolute value about the median is called minimum (you sum the absolute of everything except the median).

$$\sum_{i=0}^{n} |X_i - Median| = minimum$$

- Disadvantages of the median:

1- The median takes no account of the precise magnitude of most of the observations and is therefore less efficient than the mean
2- If two groups of data are pooled the median of the combined group cannot be expressed in terms of the medians of the two original groups but the sample mean can.

$$\bar{x}_{pooled} = \frac{n_1\bar{x}_1 + n_2\bar{x}_2}{n_1 + n_2}$$

c) The Mode:

The mode is simply the most frequently occurring value. It is possible for a distribution to have multiple modes. If all of the scores in a distribution are different, then there is no mode.
*Unstable index: values of modes tend to fluctuate from one sample to another drawn from the same population.

- 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.

-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.

- The mode may not exist.

Data: 1, 2, 3, 4, 5, 6, 7, 8, 9, 0

Here you have 10 observations and they are all different.

- The mode may not be unique.

Data: 0, 1, 1, 2, 2, 3, 3, 4, 4, 5, 5, 6, 6, 7

Mode = 1, 2, 3, 4, 5, and 6. There are six modes.

Comparison between mode, median, mean:

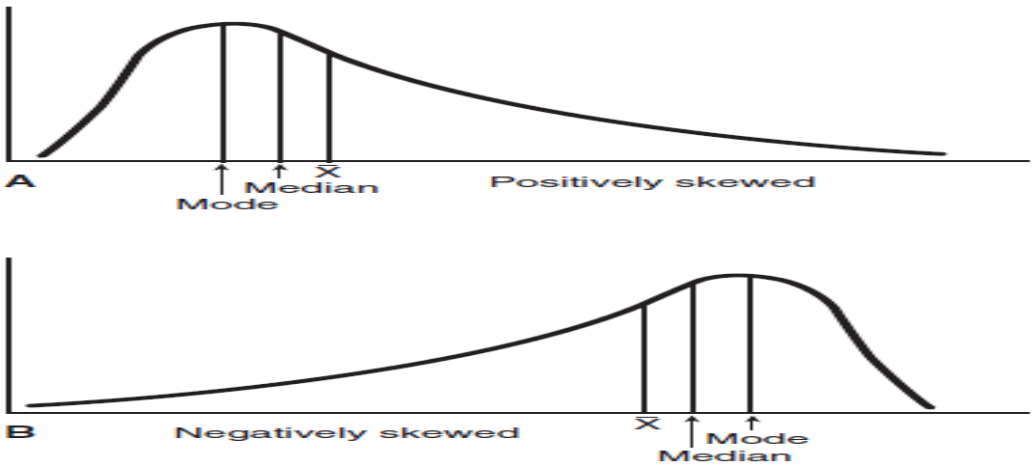
- In a normal distribution, the mode, the median, and the mean have the same value.
- The mean is the widely reported index of central tendency for variables measured on an interval and ratio scale.
- The mean takes each and every score into account.
- The mean is also the most stable index of central 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 are 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.

Summary:

- The mean is the most commonly used measure of central tendency. One

of the advantages of using the mean is that it is unique, meaning there is just one in any given data set and its calculation uses every single value in the distribution. However, the mean is also a *sensitive* measure, meaning that a single very large (or very small) value can dramatically change the mean. If a distribution is not symmetrical, the mean will not truly reflect the center of the distribution.

- The median is a more robust measure of central tendency because a few outliers or skewed distribution will not affect its value very much. As shown in the Figure, with data that are skewed to the right, the mean is larger than the median, and with data that are skewed to the left, the mean is smaller than the median.



- When a distribution is normal, meaning that it has only one mode and is symmetrical, the mean, median, and mode will have, or very nearly have, the same value. Thus, if the mean is greater than the median, then the distribution is positively skewed. If the mean is less than the median, then the distribution is negatively skewed

Good Luck