Biostatistics

Doctor 2017 | Medicine | JU

Number >>	4
Doctor	Dr Hamza
Done By	Laila Zakariya
CorrectedBy	Ghighi

*this sheet includes everything the Dr. said throughout lecture 4 with important notes from the slides written in green (slides 55-70) *

In the previous lecture, we talked about the types of variables and we ended the lecture with an important question: "Why is it important for us to know our variables and each type?" and the answer was "In order to correctly pick or select the type of inferential statistic that we need to answer our research question".

In this lecture, we will discuss how to pick an inferential statistic to answer our research question, define some concepts that we've previously mentioned in further details, and finally conclude the first chapter of this course.

Firstly, it is nearly impossible for anyone to go to every member of the population and ask them about the variables because populations are too large, usually in millions and billions. Instead, we use samples.

However, if we were able to ask each and every member of the entire population, we'll get what is called parameter.

Parameter is a descriptive measure computed from the data of the **population**.
If you want to determine the population parameters, you have to take a census of the entire population. And taking a census is very costly.
While, statistic is a descriptive measure computed from the data of the **sample**.

Example: if I want to know the mean (average) body weight of all Jordanians, I must go to all Jordanian citizens then 1) weigh them 2) add all the numbers 3) divide them by the number of Jordanian citizens (population). As a result, you'll get the populations mean of Jordanians body weight.
We cannot call this example a statistic, it is a parameter.

	Parameter (population)	Statistic (sample)
Mean	μ (mu)	$\bar{x}$ (x-bar)
Standard deviation	σ (sigma)	s
Variance	σ^2 (sigma squared)	s^2

NOTE: the Dr. drew the table above on the board and said "*for differentiating purposes only* notice how Latin characters are used in populations parameters, while English letters are used in samples statistics".

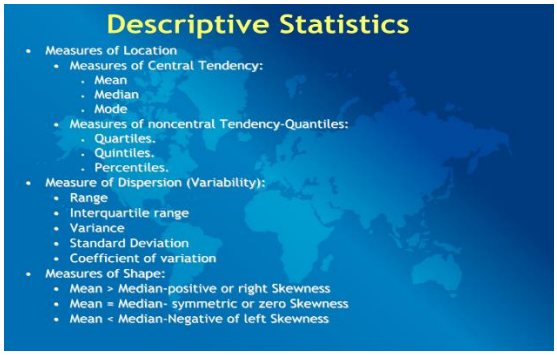
A mean of a certain value in a **population** is given the symbol μ. While, a mean of a certain value in a **sample** is given the symbol x̄, and so on.

To sum up, when we refer to a populations number, we call it a parameter and we use a Latin character to describe it. When you choose a sample because the population is too large, you express its numbers using English letters and you call it a sample statistic.

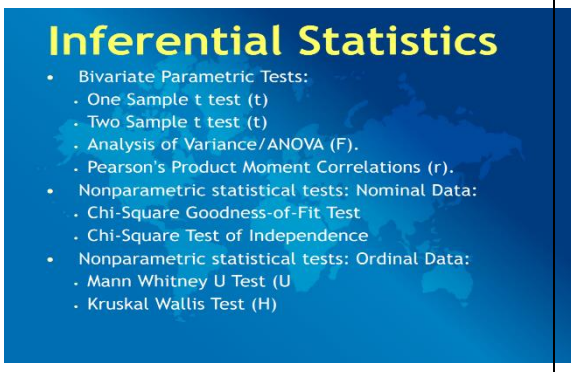
Statistics, as we know, has 2 types: **descriptive statistics** in which we try to organize, summarize, and display the numbers in a neat understandable manner.

Descriptive statistics: those statistics that summarize a sample of numerical data in terms of averages and other measures for the purpose of description, such as the mean and standard deviation.

The second type is **inferential statistics** which is the procedure used to reach a conclusion about a population based on the information derived from a sample that has been drawn from that population. So it is to predict and estimate the populations parameter.



The Dr. didn't mention anything about these 2 slides.



Example: weighing 1,000 Jordanian (a sample) to find the mean "x̄". The role and importance of inferential statistics is to predict the "μ", the mean of the population, without surveying the entire population.

The only problem with inferential statistics is that you can never be 100% confident with your results. This is because you didn't survey everybody in the population, you only surveyed a small group of people. However, you can have a certain level of confidence, and being 95% confident with your results is good enough which makes it a good predication.

What is the purpose of making predictions about the populations parameter?
To answer your research question.
Inferential statistics are used to test hypotheses (prediction) about the ***relationship***

between variables *in the population*. A relationship is a bond or association between variables.

Example: You ask if the quality of life score of Jordanian people is different than the quality of life score of Syrian refugees living in Jordan.
It is almost impossible to visit them all and ask about the quality of their lives.
The solution to this problem starts with having a hypothesis and writing it down.
For example, a hypothesis could be as the following "I guess that the mean life quality among Jordanians must be different than the mean life quality among Syrians on the population level". Meaning that there is a difference in the population parameter between population #1 (Jordanians) and population #2 (Syrians).

How can you write that in a mathematical way?

My hypothesis (H_1) is that the populations parameter mean of Jordanians (μ_J) is not the same as (not equal to ≠) the mean of Syrians (μ_S) concerning the quality of life.

$$H_1 : \mu_J \neq \mu_S$$

H1 : The researchers' hypothesis which is also known as alternative hypothesis written as H1 or HA.

H0 : The null hypothesis (the opposite of a researchers hypothesis).

If we were able to prove, using sample statistics, with a certain confidence percentage that there is a statistically significant difference. Then our hypothesis is right and we accept it.

If we didn't find a statistically significant difference, the opposite of our hypothesis would be correct. In this case, we must accept the null hypothesis, which is the opposite of our hypothesis.

NOTE: It is important to write our null hypothesis at the beginning. Why? Because at the end of our statistic, we will choose one of the two hypothesis. Either to accept the researchers hypothesis (and reject the null hypothesis) or accept the null hypothesis (and reject the researchers hypothesis). It is one of the 2 scenarios, there isn't a 3rd option.

Now back to our example, our null hypothesis would be that the mean quality of life among Jordanian citizens is not different (equal to) the mean of Syrian refugees.

$H_0 : \mu_J = \mu_S$

To make things easier, we can pick a sample of Jordanians and a sample of Syrian refugees and apply the scoring on both samples to get the $\bar{x}$ (sample mean). Then, we can apply the inferential statistics to get the μ (population mean).

Results: will be either the presence of a statistically significant difference or absence of a statistically significant difference.

Conclusion:

1) If there is a statistically significant difference ➡ I accept H1 and reject H0. Meaning that I am 95% confident that there is a difference between the quality of life among Jordanians and Syrian refugees.

OR

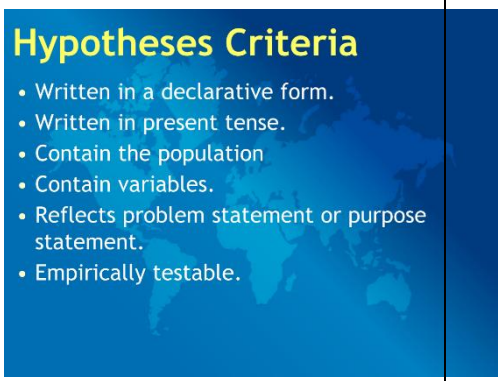
2) If there isn't a statistically significant difference ➡ I reject H1 and accept H0. And we use the word "reject H1" because there's no evidence to prove it right. In this case, we didn't reach the accepted degree (percentage) of confidence.

Technically speaking, we never accept H0. What we actually say is that we do not have the evidence to reject it.

A quick recap about hypothesis:

Research hypothesis is an **explanation of the relationship** between two or more variables and could be defined as a **translation of a research question** into a precise prediction of the expected outcomes. So it must contain terms that indicate a relationship (e.g., more than, different from, associated with).
In some way it's a proposal for solution/s.
In qualitative research, there is NO hypothesis.

This slide was not mentioned by the Dr.



A lot of mistakes and errors could occur during our testing due to many different reasons, which makes us take the wrong decision.

If you keep the null hypothesis when in fact it is right you did the right thing.

If you reject the null hypothesis when in fact it is false, you did the right thing.

However, there are 2 more possibilities that could take place:

If you reject the null hypothesis when in fact it is true (so you shouldn't have rejected it) TYPE 1 ERROR.

If you keep the null hypothesis when in fact it is false (so you shouldn't have accepted it) TYPE 2 ERROR.

The chance of committing a type I error is called α.

The chance of committing a type II error is called β.

We always want our α and β to be as little as possible.

You have to bear in mind that you cannot minimize your α and β to zero unless your sample was the whole population, and it's really hard to include the whole population in your sample. Therefore, there's an acceptable α and acceptable β. Which means accepting type I error and type II error to a certain extent. And the reason why we do this is to be reasonable (with our results).

You can't be 100% confident with your results but you can be 95% confident. This is called α.

Therefore, α is "how much you accept type I error". Accepting up to 5% is good enough so α=0.05 is the acceptable margin of type I error.

In β, you can accept up to 20% type II error. β=0.20

To sum up, you want an is 0.05 or less and β of 0.20 or less.

Why is this important? Because this is how you decide whether or not there's a statistically significant difference.

Go back to the Jordanians-Syrians example and if our α is 0.05 or less then we have to accept our hypothesis and reject the null hypothesis. However, if the final product of my calculation is above 0.05 then we must reject our hypothesis and accept the null hypothesis.

Can we lower the chances of type I error and choose an α that is less than 0.05? yes you can but the chances of getting a statistically significant difference decreases and it just makes it harder for you to find results.

That's why we try to balance between the chances of getting an error and the chances of getting a statistically significant difference.

As we lower the α error, the β error goes up: reducing the error of rejecting H0 (the error of rejection) increases the error of "Accepting" H0 when it is false (the error of acceptance).

Last but not least, the type of questions to be expected on the exam:

A researcher has rejected the null hypothesis when in fact it was false, did he:

a- Do the right thing

b- Commit a type I error

c- Commit a type II error

So you have to differentiate between type I and type II errors.

Good Luck.