2

Biostatistics

Doctor 2018 | Medicine | JU

Sheet Slides

Done By

Tasneem Jamal

Contributed in the Scientific Correction

Mohammad Abo Shaban

Contributed in the Grammatical Correction

Mohammad Abo Shaban

Doctor

Hamza Al-Duraidi

***NOTE : the underlined sentences were not mentioned by the doctor during the lecture , and when I asked him he said read everything in the slides but concentrate on what I say ☹ ***

Why we need to know the concept and the differences between variables?

Once you finish your descriptive statistics, you can now move on for inferential statistics in order to try to answer your questions (not in terms of small sample but in terms of the larger population), NOW picking the appropriate statistical test is based on your understanding of what **types of variables** you have and what are **the levels of measurements** of those variables.

1. Variables:

A variable is an object, characteristic, or property that can have different values

a **quantitative variable can be measured in some way

a **qualitative variable is characterized by its inability to be measured but it can be sorted into categories

Ex: your specialty is not a variable it's a constant because everyone in this class will provide the same answer; medicine (same observation).

-The variable is something has at least 2 different options in terms of answers like the gender (female or male).

Types of variables

In any quantitative research question, you need to have at least one independent and one dependent variable.

A. independent: the presumed cause (of a dependent variable); something cause or influence the other variable

B. dependent: the presumed effect (of an independent variable)

EXAMPLES

1. Is there a relationship between using sunblock & skin cancer?

The Independent variable: using the sunblock.

The Dependent variable: skin cancer (outcome).

2. Is there a relationship between age at marriage and happiness?

Independent: age at marriage / Dependent: Happiness.

3. Is there a relationship between smoking and lung cancer?

Independent: smoking

Dependent: developing lung cancer

(It's not because he developed lung cancer, he started smoking, NO it's because he smokes he developed lung cancer)

4. Is there a relationship between living in camps and quality of live?

Independent: living place

Dependent: quality of life

We can divide the variables into 2 main levels:

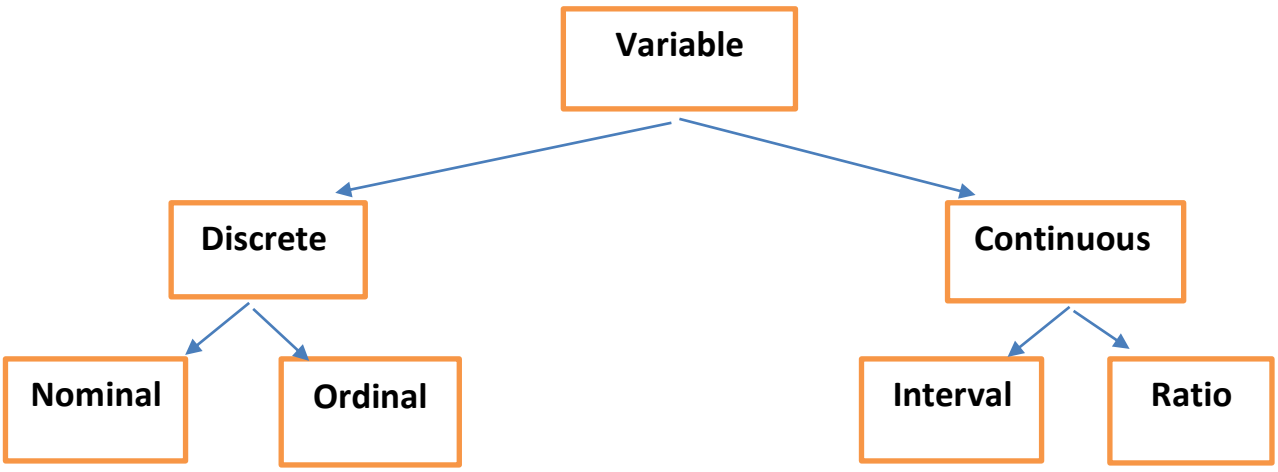
a. **CONTINUOUS**: a continuous random variable DOESN'T have gaps in the values it can assume, its properties are like the real numbers.

(Any variable that accepts fractions: age, money, height, weight, temperature, blood pressure)

b. **DISCRETE**: a discrete random variable HAS gaps or interruptions in the values that it can have, the values may be whole numbers or have spaces between them

(Any variable that does not accepts fractions: gender, color, outcome of a disease, types of food)

2. Levels of measurement:



A. Nominal:

• Categories those are distinct from each other such as gender, religion, marital status. They are symbols that have no quantitative value. Lowest level of measurement. Many characteristics can be measured on a nominal scale: race, marital status, and blood type. Dichotomous. Appropriate statistics: mode, frequency We cannot use an average. It would be meaningless here.

Ex: if you are from Amman you get 1 and if from Karak you get 2 **, here the numbers have no mathematical value**

B. Ordinal:

The exact differences between the ranks cannot be specified such as it indicates order rather than exact quantity. Involves using numbers to designate ordering on an attribute. Example: anxiety level: mild, moderate, severe. Statistics used involve frequency distributions and percentages. Appropriate statistics: same as those for nominal data, plus the median; but not the mean.

Ex: if we ask people about their educational level and we give Tawjihi…..1 , collage….. 2 , bachelor's …… 3 , PhD ……4 NOW, these numbers are meaningful because 4 is higher than 1 and so on.

When the number matters and has a mathematical value and gives an indication about certain order

(Ordinal is more powerful than nominal)

C. Interval:

They are real numbers and the difference between the ranks can be specified. Equal intervals, but no "true" zero. Involves assigning numbers that indicate both the ordering on an attribute, and the distance between score values on the attribute • They are actual numbers on a scale of measurement. Example: body temperature on the Celsius thermometer as in 36.2, 37.2 etc. means there is a difference of 1.0 degree in body temperature.

• Appropriate statistics : same as for nominal, same as for ordinal plus the mean

The zero is just another value

D. Ratio:

It's the highest level of data where data can categorized, ranked, difference between ranks can be specified and a true or natural zero point can be identified. A zero point means that there is a total absence of the quantity being measured. All scales, whether they measure weight in kilograms or pounds, start at 0. The 0 means something and is not arbitrary (SUBJECTIVE). Example: total amount of money.

When you say zero you mean zero; the absence of the variable you are measuring

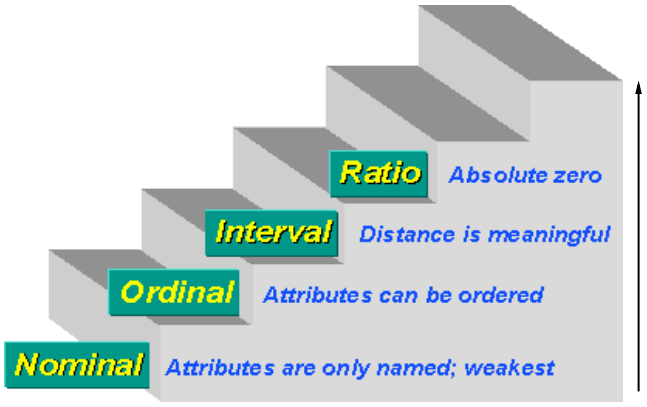
**EXAMPLES **

1. If the **temperature is zero** that does not mean the absence of temperature, so it is an **interval** because the zero in it does not mean the absence of the variable

2. If you have **zero money** that means the absence of money, so it is a **ratio**

3. How much **mercury** there is in the blood sample?

If it is zero that means there isn't any mercury in this sample, **and you can't say there is (-2) because this is a ratio and it begins at zero and the zero in it means an absolute zero (there is no acceptable value below zero but we accept fractions on it)**



Wink wink:

Ratio is the most powerful level of measurements.

What type of data to collect?

The goal of the researcher is to use the highest level of measurement possible.

EX: two ways of asking about smoking behavior .which is better A or B?

A. do you smoke? yes or no

B. how many cigarettes did you smoke in the last 3 days (72 hours) ?

(A) is nominal , so the best we can get from this data are frequencies . (B) is ratio , so we can compute : mean, median, mode, frequencies .

Parameter and Statistic

The term **parameter** is used when describing the characteristics of **the population**. The term **statistics** is used to describe the characteristics of **the sample**.

Types of Statistics:

• **Descriptive Statistics**: It involves organizing, summarizing & displaying data to make them more understandable, those statistics 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.

• **Inferential Statistics**: It reports the degree of confidence of the sample statistic that predicts the value of the population parameter.

They are used to test hypotheses (prediction) about relationship btw variables in the population.

**statistical inference: is the procedure used to reach a conclusion about a population based on the information derived from a sample that has been drawn from the population.

Dependent vs. independent & continuous vs. discrete these two things are very important to understand what types of inferential statistics you will need to answer your research question:

*if you have a **dependent** variable that is **continuous** : you use parametric inferential statistics like t-test or ANOVA .

*if you have **dependent** variable that is **discrete** : you use nonparametric inferential statistics like Chi-square.

NOW we know that we are going to use parametric but which one of the parametric tests?

We go to the **independent** variable, if it's:

1- Continuous: Pearson's product moment correlations (r).

2- Discrete (with 2 values only): t-test

3- Discrete (with more than 2 values) : ANOVA

****BUT don't worry, in this course we are just going to take one example on parametric (t-test) and one on nonparametric (Chi-square)****

****

At the end of any quantitative research the decision (conclusion) is one of two things:

1. Yes, we find statistically significant association between the 2 variables so we reject the null hypothesis

2. We are not able to find statistically significant association between the 2 variables so we keep the null hypothesis

The null hypothesis: it is the negative hypothesis that is saying there is no association between variables.

EX: when a research idea come to our minds we have a hypothesis that we are trying to proof (like if there is a relationship btw using sunblock and skin cancer), this is called the researcher hypothesis and the opposite of it that there is no relationship is called the null hypothesis.

And our aim is always to reject it and proof our hypothesis.

For many reasons researchers sometimes commit mistakes (maybe they didn't choose the sample correctly or didn't use the correct inferential test,…..etc) so , there is 2 types of errors in the quantitative research :

1) Type 1 error (alpha…α) **2) Type 2 error (beta….β)**

this scenario will clarify everything

We want to measure the association btw using sunblock and developing skin cancer, so we could end up with one of 4 decisions:

1) Reject the null hypothesis (I found statistically significant association btw the 2 variables when **in fact the null hypothesis was true**)→ **type 1 error**

2) Reject the null hypothesis (I found statistically significant association when in fact you were right; **the null hypothesis was false**)→ **no mistake**

3) You were not able to find a statistically significant association so you kept the null hypothesis when in fact **it was false you should have rejected it** → **type 2 error**

4) you kept the null hypothesis when in fact it was really true → no mistake

**we can't be absolutely sure if we have done the right or wrong thing , but the beauty of statistics is that we are able to control to what extent we allow each type of error (the maximum acceptable margin of error)

Type 1 error: we can only allow up to 5% chance that this error happens.

α → 5%

(1-α) = 95% (**confidence**)

> We have decided from the beginning that we want our level of confidence to be 95% because we want to allow up to 5% chance for type 1 error.

Type 2 error: it is less dangerous so we all a margin (chance) of error up to 20%.

β→20%

(1-β) = 80% (**power**)

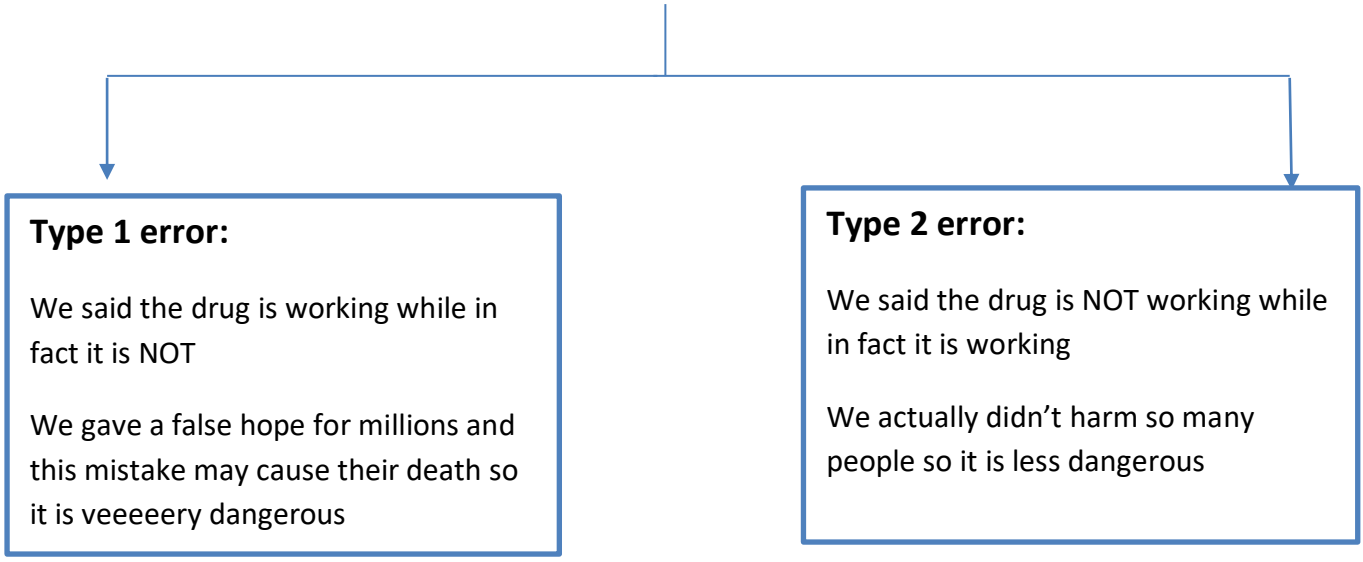
> If we want 100% power and confidence we will end up collecting data from everyone in the population which is impossible so we allow the 5% and 20% to be flexible with the size of our sample .

In our conclusion, we don't say for example that using sunblock leads for sure to skin cancer because we acknowledge that we collect data from small sample , so we say that we are 95% confident about the result .

*In physics, chemistry , genetics ,etc. 5% is a huge number for chance of error so their α is (0.001 or maximum 0.01) ; because it is a highly controlled environment . while in studying human behavior and human health , it's a very complex thing and there are many factors contributing in (most of them we can't consider) so we can accept an α of 5% .

*** Type 1 error is more dangerous and harmful than type 2 error**, why?** ***

EX: If there is a poor African country where AIDS is killing hundreds of people daily and the drug is very expensive that even the government can't buy, now if someone made a new drug that can cure AIDS and at the same time it's much less expensive → we take a sample of people and try the drug on them to make sure it is working.



Q: what is the difference btw confidence and power?

Confidence: to what limit we are sure that the result we came with is based on a real association btw the variables

Power: to what extent your sample size and statistical test were powerful to detecting the association btw variables

Wink wink: it is known that α is 5% BUT it could be different

So before answering any question in biostatistics take a careful look at the question if the value of α is different or not , and at different levels of α you take different decisions regarding rejecting or keeping the null hypothesis