◉Sheet ○Slides

Number

4

Done by:

Ola Al-juneidi

corrected by:

Basheq Jehad

Doctor

Mahmoud Al-Hussami

Record #5 Slide #1-slides (40-57)

Review Question:

*What is the best type of variable that can be collected?

- The continuous type is better than any other type.

EX: If we ask someone: "do you smoke?" He will answer with YES or NO (Dichotomous variable). Now to make it continuous we ask "how many cigarettes do you smoke every day?" The answers will vary (0, 10, 20 etc.). Here we can deal with data with all types of statistics (we can compute the mean, median, mode, frequencies..)

Parameter and Statistic

Parameter: is a descriptive measure computed from the data of the *population*.

- The population mean, μ, and the population standard deviation, σ, are two examples of population parameters.
- Since the population is not actually observed, the parameters are considered unknown constants.

Statistic: is a descriptive measure computed from the data of the *sample*.

- For example, the sample mean, $\bar{x}$, and the standard deviation, s, are statistics.
- They are used to estimate the population parameters.

So to generalize for the general population we use parameter, but it is hard to study all the characteristics for a population since taking a census is very costly; it needs money and effort. We can take a sample and generalize the results on the whole population.

The techniques used to test a hypothesis are called **parametric techniques**.
*Why did we call them PARAMETRIC not STATISTICAL? Because they are related to the population. We use them to generalize it –the hypothesis- on the population with a possibility of error (the standard error of the mean-we'll talk about it later) which we must take into consideration.

Statistics is a branch of applied mathematics that deals with collecting, organizing, & interpreting data using well-defined procedures (parametric techniques) in order to make decisions.

Statistics are different from BIOstatistics; Statistics deal with every single thing, but biostatistics deal with health data related to humans.

Types of Statistics:

1. **Descriptive Statistics**: It involves organizing, summarizing & displaying data to make them more understandable. It includes the presentation of data in the form of graphs, charts, and tables so that we can read them easily.

   - Most of the times we don't use descriptive statistics to generalize. In other words, descriptive statistics are not concerned with the theory and methodology for drawing inferences that extend beyond the particular set of data examined - from the sample to the entire population - . All that we care about are the summary measurements, such as the average (mean).
   - EX: When we calculate the mean, median, and the distribution of marks after a midterm exam, we don't want to generalize anything on the students. Instead, we only want to describe and summarize the data.

   - **Types of descriptive statistics:**
     - **A. <u>Measures of Location:</u>**

       <u>A.1</u> Measures of Central Tendency: mean, median, mode.

       <u>A.2</u> Measures of noncentral Tendency (Quantiles): Quartiles, Quintiles, and Percentiles.

       EX: When dealing with marks we usually use quartiles (dividing students into 4 quartiles), for the lowest quartile (25% of students) we give the marks C and below and then continue for the rest. But if we used the median, half of the students will get C+ and below! We also don't use the mean because it is sensitive to extreme values.

     - **B. <u>Measure of Dispersion (Variability):</u>**

       <u>B.1</u> Range

       <u>B.2</u> Interquartile range

       <u>B.3</u> Variance

       <u>B.4</u> Standard Deviation

       <u>B.5</u> Coefficient of variation

Standard deviation and coefficient of variation are basically the same. We use standard deviation to compare things of the same type, while coefficient of variation is used for the comparison of different things. EX: If we compare two types of gold we use standard deviation, but if we compare one type of gold with one type of silver we use coefficient of variation.

C. **<u>Measures of Shape:</u>**
   - Mean > Median-positive or right Skewness
   - Mean = Median-symmetric or zero Skewness
   - Mean < Median-Negative of left Skewness

*We will talk about each one of them more in the upcoming lectures.

00:00-16:00

Slides 40-44

2. **Inferential Statistics**: It reports the degree of confidence or accuracy of the sample statistic that predicts the value of the population parameter. We do this by using procedures (either parametric or non-parametric) to reach a conclusion (or to generalize) about that population based on the information derived from the sample that has been drawn from that population, or to test a hypothesis about the relationship between variables in the population. A relationship is a bond or association between variables.
   - For us to be able to generalize we must pick a random sample and every single subject in the population has to be selected.

   - **Types of inferential statistics:**

     <u>1.Bivariate Parametric Tests:</u>

     (a) One Sample t test (t)
     (b) Two Sample t test (t)
     (c) Analysis of Variance/ANOVA (F)
     (d) Pearson's Product Moment Correlations (r) >> we use it if the two variables are continuous (intervals and ratios).

     $-1 \leq r \leq 1$ where: 1 means that the relationship between the two variables is a total positive linear correlation.
     -1 means that it's a total negative linear correlation.

0 means that there is NO correlation between the 2 variables.
If it's + >> positive correlation
If it's - >> negative correlation

If r is: below 0.6 → the relation is weak

0.6-0.79 → moderate relation

0.8 and above → strong relation

*Question: Which is bigger r= -0.9 or +0.5?
- The correct answer is -0.9, because we look at the absolute value of the number. (+) or (–) only determines the DIRECTION not the value.

EX: the correlation between the number of lectures you attend and your grade is a POSITIVE correlation, while the correlation between the number of lectures you were absent and your grades is a NEGATIVE correlation.

Note: Univariate tests are used in *descriptive* statistics because they are concerned with only one variable. In bivariate tests we study the relation between 2 variables (ex: the effect of X on Y, and that of J on Y). We are not concerned with multivariate tests in this course.

2. Nonparametric statistical tests: Nominal Data (1 group or more):
   (a) Chi-Square Goodness-of-Fit Test
   (b) Chi-Square Test of Independence
   We choose between (a) or (b) according to the groups; being dependent or independent:
   - If it's 1 group we use pre and post test.
   - If there are 2 or more groups we have 2 options:
     -for dependent groups >> (a)
     -for independent groups >> (b)

3. Nonparametric statistical tests: Ordinal Data (2 groups or more):
   (a) Mann Whitney U Test (U) >> for 2 groups
   (b) Kruskal Wallis Test (H) >> for more than 2 groups

→For continuous data:

a) For 1 group:
   - Z-score
   - T-test
   - 2 means (pre & post) >> dependent t-test

b) For 2 groups
   - 2 means >> independent t-test

c) 2 or more groups:
   - One way ANOVA

→Tests for correlations between variables:

- Both variables are continuous >> Pearson's correlations (as explained earlier)
- Both variables are ordinal data >> Spearman test
- If the dependent variable was continuous and the independent was categorical/nominal >> t-test or z-test
- If the dependent variable was continuous and the independent was ordinal >> point biserial correlation

*We will talk more about these different tests and when to use each one of them.

16:00-27:00

Slides 45-47

When studying the correlation between different variables we're not studying the effects. Correlational studies come after *observational studies* and before *clinical trials*; they come in the 2nd stage. We study correlations before studying the cause-effect relationship, and only then we proceed to more advanced designs in research. If there was no correlation between the variables, we don't proceed to the advanced steps.

EX: Studying the relation between smoking and lung cancer. First, they studied it by *descriptive studies* and explorations; they noticed some connection. Then, they studied the *correlation* and the strength of this correlation between the two; they found a strong relation.

When they found that, they moved to *cause-and-effect studies*; does smoking LEAD to lung cancer or not? Here we're talking about more advanced designs (*analytical designs*), and by clinical trials, they found that smoking is a CAUSE of lung cancer.

27:00-36:00

Research hypothesis

A **hypothesis** is made about the value of a parameter, but the only facts available to estimate the true parameter are those provided by the sample. If the statistic differs (and of course it will) from the hypothesis stated about the parameter, a decision must be made as to whether or not this difference is significant. If it is, the hypothesis is rejected. If not, it cannot be rejected.

There are 2 types of hypotheses:

1. **Null hypothesis** or statistical hypothesis "الفرضية الصفرية" (H0): This contains the hypothesized parameter value which will be compared with the sample value. It is used in statistics. We assume that there is no relation between variables and we want to approve the opposite.

2. **Alternative hypothesis** or research hypothesis "الفرضية البحثية" (H1): This will be "accepted" only if H0 is rejected.

EXAMPLE: Studying the relationship between attendance and grades of students. "There IS a positive relationship between the presence of students in class and their mid-term grades" >> this is the **alternative hypothesis**, it is divided into 2 parts: directional or non-directional.

"There is no relationship between the presence of students in class and their mid-term grades" >> this is the **null hypothesis**

Technically speaking, we never accept H0. What we actually mean is that we do not have the evidence to reject it. So our goal now is to prove that the null hypothesis is WRONG, but how?

Steps in hypothesis testing:

1. First, we collect data: your attendance and grade.

2. Write the null hypothesis (H0) down.

3. We have to establish the α value. It can be:
   a) 0.05 b)0.01 c) 0.001

*But what is the α value?

The researcher usually looks for the effect of the independent variable on the dependent variable and the result must be coming totally from the independent variable and not from something else. Actually, there is no such thing as 100% in research (i.e. X caused Y 100% or Y is a result of X and only X). There is always something coming by chance (not from the studied variable) i.e. related to the extraneous variables or confounding variables or any other unknown variable.

According to this, we define α (Significance level) as the probability of rejecting the null hypothesis when it is, in fact, true. It is the probability of the results that came by chance.

*The largest value of α is 0.001 and the smallest is 0.05, HOW??

-Because we look at the probability of the results that DIDN'T come by chance but came from the independent variable, that is 0.999 in the case of 0.001 and 0.95 in the case of 0.05. NOW we compare: 0.999 > 0.95, so 0.001 is larger than 0.05 (as an α value).

*How do we choose the value of α?

-In behavioral science we choose it to be 0.05, because it is related to behaviors and it's not really important so we choose it to be the least. However, if our study is related to drugs or any other sensitive subject we choose it to be 0.01 or 0.001 according to the importance of the subject.

4. Establish the ***critical value*** or values of the test statistic needed to reject H0.

5. Select the test statistic (as said before): t, Z, F, Chi-square…. we call this the ***calculated value***. Whenever the calculated value increases, the t value decreases.

6. Make a decision: Reject H0 or don't reject H0. If the calculated value turns out to be less than the cut-off (α) value that we determined we reject the null hypothesis. If it's more than the cut-off value we cannot reject the null hypothesis. (Know this for now, but we will clarify it more later on).

Types of errors while testing hypotheses:

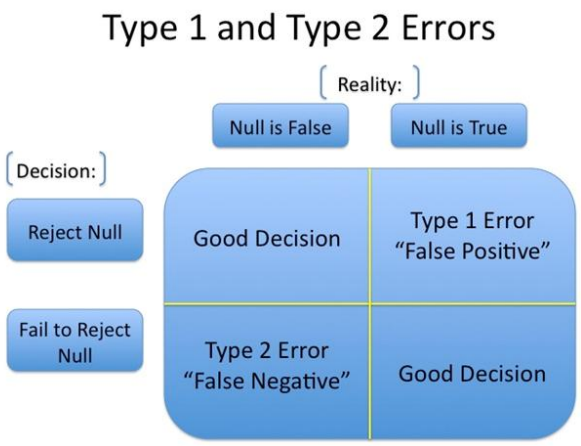
α (or type 1) error:

- You make this type of error when you reject the null hypothesis while it's true.
- Often committed by novice researchers and practitioners.

β (or type 2) error:

- You make this type of error when you accept the null hypothesis while it's false.
- Often committed by advanced researchers.

If You……	**When the Null Hypothesis is…**	**Then You Have…….**
Reject the null hypothesis	True (there really is no difference)	Made a Type I Error
Reject the null hypothesis	False (there really is a difference)	☺
Accept the null hypothesis	False (there really is a difference)	Made a Type II Error
Accept the null hypothesis	True (there really is no difference)	☺



TRADEOFF!

- There is a tradeoff between the alpha and beta errors. We cannot simply reduce both types of error. As one goes down, the other rises.
- 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).

EXAMPLE: Quality Control.

- A company purchases chips for its smart phones in batches of 50,000. The company is willing to live with a few defects per 50,000 chips. How many defects?
- If the firm randomly samples 100 chips from each batch of 50,000 and rejects the entire shipment if there are ANY defects, it may end up rejecting too many shipments (error of rejection). If the firm is too liberal in what it accepts and assumes everything is "sampling error" it is likely to make the error of acceptance.
- This is why government and industry generally work with an alpha error of 0.05.

36:00-45:00

Slides 48-57

"Never give up on a dream just because of the time needed to accomplish it. The time will pass anyway"

GOOD LUCK