Public health

Sheet #4 Abdelaziz Al-Shawa

The Dr showed a data sheet (input sheet) on SPSS system. Nowadays with such software’s, we don’t need to know the mathematical equations, we only need to know the tests we want to conduct and the software will do the job, this software can do all types of statistical analysis or measures. We get the results as (output sheet)

In SPSS we have 3 sheets : 1. Input sheet , 2. Variable view sheet , 3. Output sheet, we can study our data through any of these sheets.

When the value of ; mean , mode , median are close together ,this indicates that the advanced statistical analysis (inferential statistics) can be done correctly on this data; the closer ,the better.

Reminder: “[*Univariate analysis*](https://en.wikipedia.org/wiki/Univariate_analysis)*involves describing the*[*distribution*](https://en.wikipedia.org/wiki/Frequency_distribution)*of a single variable, including its central tendency (including the*[*mean*](https://en.wikipedia.org/wiki/Mean)*,*[*median*](https://en.wikipedia.org/wiki/Median)*, and*[*mode*](https://en.wikipedia.org/wiki/Mode_(statistics))*) and dispersion (including the*[*range*](https://en.wikipedia.org/wiki/Range_(statistics))*and*[*quantiles*](https://en.wikipedia.org/wiki/Quantiles)*of the data-set, and measures of spread such as the*[*variance*](https://en.wikipedia.org/wiki/Variance)*and*[*standard deviation*](https://en.wikipedia.org/wiki/Standard_deviation)*). The shape of the distribution may also be described via indices such as*[*skewness*](https://en.wikipedia.org/wiki/Skewness)*and*[*kurtosis*](https://en.wikipedia.org/wiki/Kurtosis)*. Characteristics of a variable's distribution may also be depicted in graphical or tabular format, including*[*histograms*](https://en.wikipedia.org/wiki/Histograms)*and*[*stem-and-leaf display*](https://en.wikipedia.org/wiki/Stem-and-leaf_display).”

-***The shape of the distribution*** provides information about the central tendency and variability of measurements.

Three common shapes of distribution are:

**-Normal** distribution (bell shaped)

-**skewed**: piled up to the left (positive skewness) or piled up to the right (negative skewness)

\*Usually It’s easier to determine the skewness from the equation rather than the chart

-**Multimodal:** (bimodal or trimodal) has more than one peak (mode) , it’s not a normal distribution but if the bimodal (for example) has two close modes that are close to mean and median , we can do assumption ; and do inferential statistics on such data.

**Skewness**

-It is the measure of the shape of a nonsymmetrical distribution

in skewed data (positively or negatively). there’s a cutoff point at which data can’t be suitable to do inferential statistics on it, so you should correct the data first. This correction of data is called ***Data Transformation***, which means: to transform/correct skewed data to make distribution appear more normal or symmetrical so you can apply inferential statistics.

There’re two equations to measure skewness:

* **Pearson’s Skewness Coefficient Formula = (mean-median) / SD**

The cutoff point for this formula is 0.2, We take the absolute value of the result to compare it with the cutoff point, the + or – sign indicates just the direction of the skewness , negative or positive skewness

\*Skewness values > 0.2 or < 0.2 indicate severe skewness

* **Fisher’s Skewness Coefficient Formula** (more accurate than Pearson’s) **=**

**Skewness coefficient NB / Standard error of skewness**

**\*** Skewness coefficient NB**;** that was calculated in Peterson’s formula

\*\*standard error of skewness: is a standard value calculated for overall population among certain variable (age for example) , so it’s value will be given in the question

The cutoff point is 1.96 SD (*Skewness values >+1.96 SD* indicate severe skewness) , also here we take the absolute value of the result to compare it with this cutoff point.

-*Standard error* vs *standard deviation*: standard deviation it’s a measurement for the sample but the standard error is used for the whole population; it’s formula is ( SD /) while N is the population , this equation is not required for the exam, just for better understanding , and the standard error is usually calculated using SPSS.

*The first assumption to do parametric statistics is* ***normality*** *(proximity to normal distribution) and it’s tested by skewness*

**Inferential statistics**

Statistical inference : is the only way to take a decision about the **relation** between two variables or more (ex. Is there a relation between lung cancer or smoking or not? , you can’t decide just from the prevalence of lung cancer in smokers and non-smokers , you have to do inferential statistics.

The statistical inference is of two types: 1. *parameter estimation* and

2. *hypothesis testing: which is usually used in the medical field and consists of* ***5 steps*** ,

after that you can decide if there’s a real relation between these two variables or not “to accept or reject the hypothesis”)

The explanation of inferential statistics will be through this example:

Ex. A study in University of Jordan hospital about Hematoma post breast reduction (mammoplasty);

66 ladies did mammoplasty, some of them had hematoma after the operation and some didn’t and they have studied many risk factors to see if there’s a relation or not of each factor/variable with hematoma formation:

Many possible risk factors have been studied, as: Anti-coagulants intake (heparin), smoking, medical illness, Age

We can not decide just from the frequency: 6 ladies was taking anticoagulants and had hematoma, 31 ladies was taking anticoagulants and didn’t have hematoma, we can’t decide just from these numbers if there’s a relationship or not.

To decide if there’s an association between anticoagulants intake and hematoma, we must do statistical inference and observe the probability ( p-value )

If the probability was equal to or less than alfa ( α ); so there’s a significant relationship

If the probability was more than alfa ( α ); so there’s no significant relationship

**The steps of hypothesis testing :**

***The First step*** to examine any statistical relationship between two variables using statistical inference is to set the hypothesis

Between the variables of interest:

Hypothesis: is a tentative prediction of the researcher about a relationship between two or more variables

-Two types of hypotheses:

*Research hypothesis/alternative hypothesis H1*: there’s statistically significant relationship between the variables of interest, anticoagulant intake and hematoma development post mammoplasty in our example

*Null hypothesis H0* : there’s **NO** relationship between the variables of interest, anticoagulant intake and hematoma development post mammoplasty in our example

-In statistics and scientific research we have to prove/disapprove the **null** hypothesis

If you accept the null , you have to reject the research hypothesis

If you reject the null , you have to accept the research hypothesis

You can’t accept or reject both of them.

***The Second step*** is to set a risk level (α): which is the degree of risk of incorrectly concluding Ho is false when it is true, it can be 0.05 or less.

-if you are doing (descriptive statistic/non advanced statistics/modest research/not very controlled research) , all scientific research textbooks say that α should be set at 0.05

When I set α at 0.05 this means that I assume that the possibility of the absence of a relationship between that two variables of interest is 0.05 , so I’m confident by 0.95 (the complement of 0.05) that there’s a relationship between them.

Risk level (0.05) + confidence level (0.95) = 1.00

So the risk level should be 0.05 or less (0.00001 for example) this depends on the experience of the researcher and the number of times this study have been done before

\*The risk level can never be more than 0.05

So in exam questions you always assume α = 0.05 unless noted otherwise

***The Third step*** is to choose the appropriate statistical test (parametric/nonparametric) to test the null hypothesis. \*you choose the test you need depending on the nature of data you have

In this example they choose *fisher’s exact test* which is one type of *chi square* tests “which is non parametric” (they chose to do non parametric statistics because the two variables , the dependent which is the Hematoma and the independent which is the anticoagulation, both of them, are at nominal level so we can use chi square test which is suitable in such a case)

***The Fourth step*** is to calculate/observe probability using the test you’ve chosen (using a software), so you get the possibility of occurrence/non-occurence of that phenomenon

In this example The observed probability is 0.031 which means that possibility of **not having** a relationship between anticoagulant taking and hematoma development is 0.031 according to this sample

***The Fifth step*** is to accept or to reject the null hypothesis after comparing the observed probability (p-value) with the risk level (α), (if p is less than or equals α, you have to reject the null hypothesis)

So in this example 0.031 is less than 0.05 so we reject the null hypothesis and accept the alternative one ; which means : there is a statistically significant relationship between anticoagulants and hematoma formation.

* Another example from the same study:

the variable of interest is the Age , and we need to decide if the age of a lady plays a role in hematoma formation or not

1. Set the null hypothesis : there’s no relationship between age and hematoma formation post mammoplasty
2. Risk level : α = 0.05
3. Statistical test diagnostically used: they used *Student’s t test* (parametric statistical test that compares two means)
4. Observed p-value : after execution of of *Student’s t test* we get the observed probability as a result, which is here 0.175
5. The final decision (compare p-value with α) : 0.175 is more than 0.05 so we must accept the null hypothesis which states that there is No relationship between age and hematoma formation

**Don’t forget that every expert was once a beginner :)**

**Best wishes**