Quantitative Methods Research in Language and Linguistics & SPSS: An introduction

Ángela Almela, PhD
Universidad de Murcia

November 8, 2021

Index

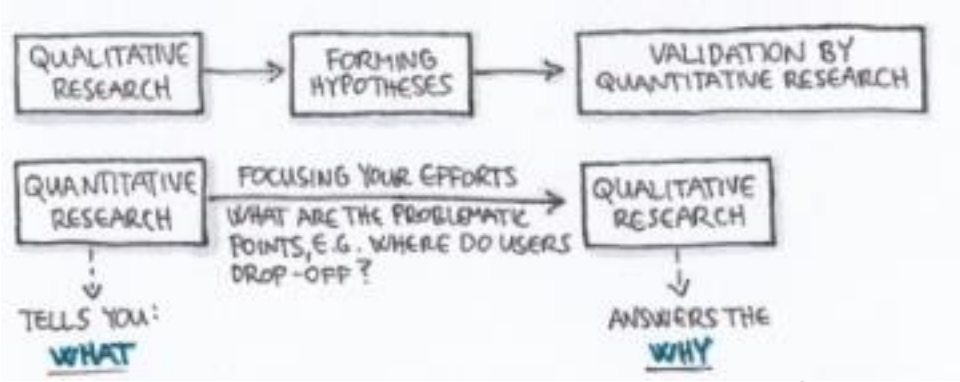
- Why using statistics in linguistic research?
- Some basic notions: descriptive statistics
  - Central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability or dispersion
    - Variance
    - Standard deviation
- Scales and variables
- Basic notions of inferential statistics

Index

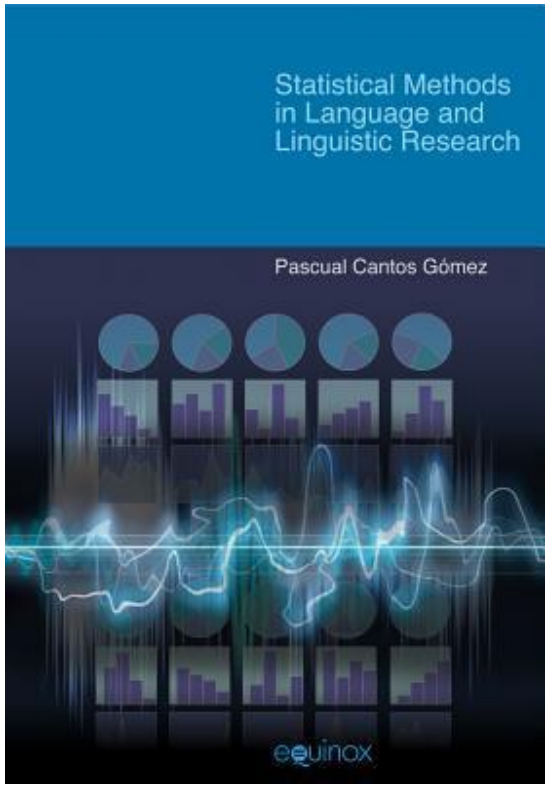
- Why using statistics in linguistic research?
- Some basic notions: descriptive statistics
  - Central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability or dispersion
    - Variance
    - Standard deviation
- Scales and variables
- Basic notions of inferential statistics

Why using statistics in linguistic research?

- Qualitative research
  - Use of data to identify and describe features of language usage
  - To provide real occurrences of particular phenomena
- In quantitative analysis...
  - Linguistic features are classified and counted
  - (Complex) statistical models are constructed so as to explain the observed facts



Reference work
Cantos (2013)



Statistics: some basic concepts

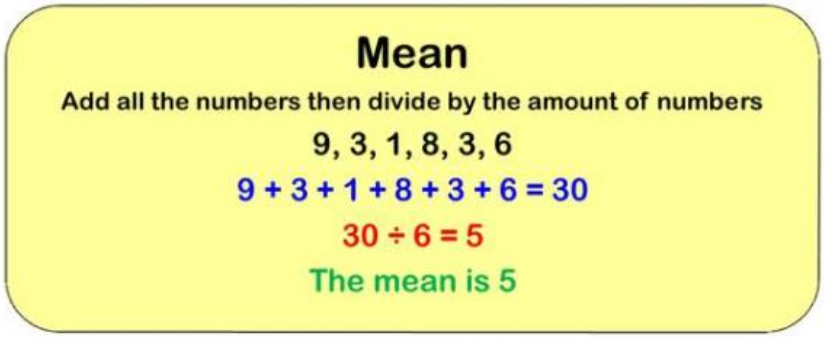
- Population: total universe of all possible observations.
  - e.g. The weight of all adult women in Spain
  - Descriptive statistics are called *parameters*
- Sample: subset of observations drawn from a given population
  - The weight of 100 adult women in Spain
  - Descriptive statistics from a sample are called *statistics*
  - It is important to draw a representative, random sample that is not biased

Index

- Why using statistics in linguistic research?
- Some basic notions: descriptive statistics
  - Central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability or dispersion
    - Variance
    - Standard deviation
- Scales and variables
- Basic notions of inferential statistics

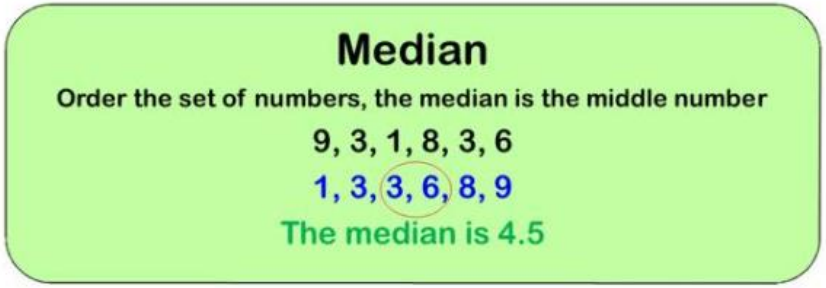
Central tendency: mean

- Mean: sum of scores divided by the number of scores
  - Every score is taken into account
  - Some interesting properties ➔ Most widely used



Central tendency: median

- Median: score from the middle of the list when ordered from lowest to highest
  - Cuts data into halves, not taking into account values of all scores but only the scores in middle position
  - Good estimate for abnormally large or small values
  - Only influenced by values in the middle of ordered data



Central tendency: how to calculate the median

$\mathbf{med(x)} = \left(\frac{n+1}{2}\right)^{\text{th}}$ smallest value if n is odd

$\mathbf{med(x)}$ = average of the $\left(\frac{n}{2}\right)^{\text{th}}$ and $\left(\frac{n}{2}+1\right)^{\text{th}}$ smallest values if n is even

Caution: The observations must be arranged first! (preferably in an *ascending order*)

Central tendency: mode

- Mode: Most frequent mark (Note: there may be multiple modes)
  - Quick if we have frequency distribution
  - Possible with *categorical* data

frequency distribution table

A data table that lists a set of scores and their frequency.

score	tally	frequency (f)
1	𝍸	4
2	𝍸 𝍸	9
3	𝍸	6
4	𝍸	7
5	𝍸	3
6	𝍸	2

Mode

The most common number

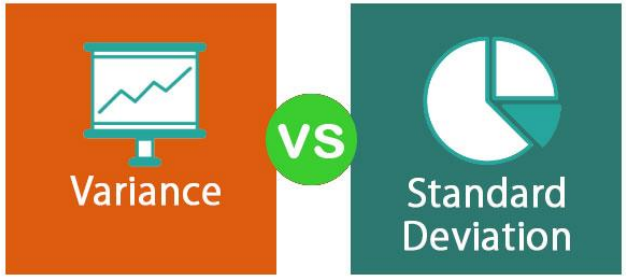
9, 3, 1, 8, 3, 6

The mode is 3

Index

- Why using statistics in linguistic research?
- Some basic notions: descriptive statistics
  - Central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability or dispersion
    - Variance
    - Standard deviation
- Scales and variables
- Basic notions of inferential statistics

Spread of the population: variability measures



- Variance: sum of squared deviations from the mean
  - Used to obtain the measure between the variables and how they differ from one another
- Standard deviation: square root of variance
  - It shows how the observations/values differ from the mean or the average value from the data set

Normal distribution

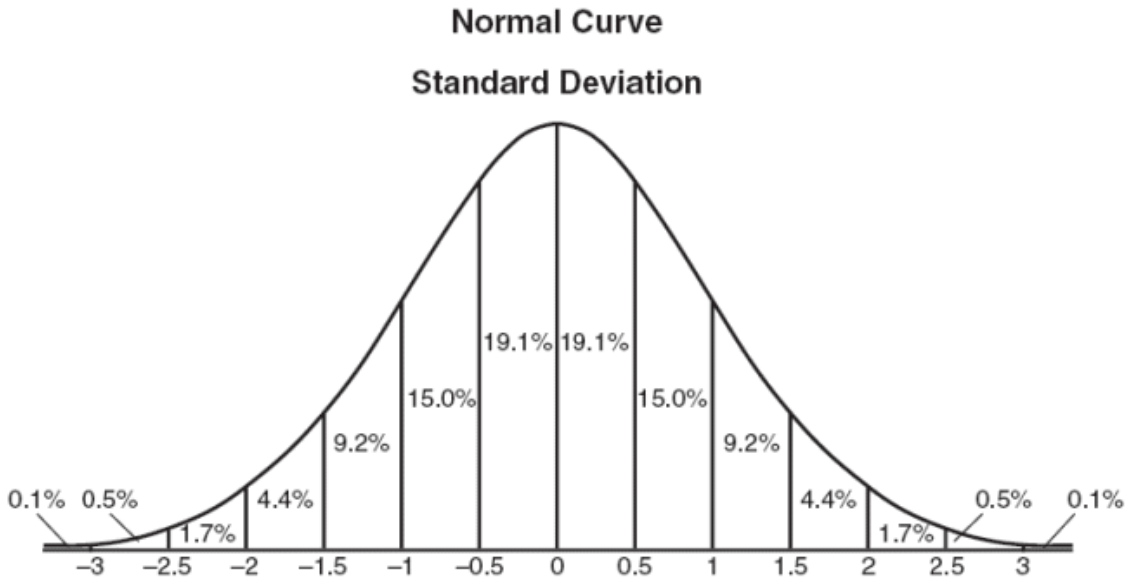


Chart prepared by the NY State Education Department
(http://www.regentsprep.org/regents/math/algtrig/ats2/normallesson.htm)

The spread of a normal distribution is controlled by the standard deviation. The smaller the standard deviation, the more concentrated the data.

Spread of the population: standard deviation

Sentence length (in words)	Domain	Medium
12	Mathematics	oral
12	Literature	written
13	History	oral
14	Mathematics	oral
15	Mathematics	written
16	History	written
16	Literature	written
17	Literature	oral
20	History	written
22	Literature	written

Spread of the population: standard deviation

Medium	N	Mean	Standard Deviation
Oral	4	14	2.16
Written	6	16.83	3.6
Total	10	15.7	3.3

It shows how the observations differ from the mean or the average value from the data set

Don’t panic about all these formulae…

- Some good news! Fortunately, we have spreadsheets and software packages like SPSS to obtain all these measures automatically.

$$V = \frac{\sum diff^2}{N-1}$$

$$\bar{X} = \frac{\sum x}{N}$$

$$\sqrt{\frac{\sum (x-\bar{x})^2}{(n-1)}}$$

Index

- Why using statistics in linguistic research?
- Some basic notions: descriptive statistics
  - Central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability or dispersion
    - Variance
    - Standard deviation
- Scales and variables
- Basic notions of inferential statistics

TYPES OF SCALES

- Nominal scales
- Ordinal scales
- Interval scales
- Rational scales

TYPES OF SCALES

- NOMINAL OR CATEGORICAL SCALE
  - Nominal is hardly measurement (categorical or qualitative variables)
  - Quality more than quantity: categorization of items
  - Nominal level of measurement is a matter of distinguishing by name, e.g.
    - 1 = male, 2 = female
    - Binary category of 0 and 1 used for computers
    - Meal preference: Breakfast, Lunch, Dinner
    - Religious preference: 1 = Buddhist, 2 = Muslim, 3 = Christian, 4 = Jewish, 5 = Other
    - Political orientation: Republican, Democratic, Libertarian, Green

TYPES OF SCALES

- ORDINAL SCALE
  - Ordinal refers to order in measurement - used to rank data
  - Ordinal scale = nominal information + direction
    - Ranking an experience as a 6 on a scale of 1 to 10 tells us that it was higher than an experience ranked as a 3, but with no implication of twice.
    - Many psychological scales or inventories are at the ordinal level of measurement.
  - Examples:
    - Low/Medium/High
    - Faster/Slower
    - Rank: 1st place, 2nd place, ... last place
    - Level of agreement: No, Maybe, Yes
    - Political orientation: Left, Center, Right

TYPES OF SCALES

- INTERVAL SCALE
  - Interval scales provide information about order, and also possess equal intervals.
  - If we can state that the distance between 2 and 3 is the same as that between 6 and 7 on a 10-point rating scale, then we have an interval scale.
  - Temperature, either measured on a Fahrenheit or Celsius scale
    - A degree represents the same underlying amount of heat, regardless of where it occurs on the scale. Measured in Fahrenheit units, the difference between a temperature of 36 and 32 is the same as the difference between 84 and 80.
  - Equal-interval scales of measurement can be devised for opinions and attitudes.
  - Further examples:
    - Time of day on a 12-hour clock
    - Other scales constructed so as to possess equal intervals

TYPES OF SCALES

- RATIONAL SCALE
  - Qualities of nominal, ordinal, and interval scales + an absolute zero (a point where none of the quality being measured exists)
  - Using a ratio scale permits comparisons such as being twice as high, or one-half as much.
  - Some examples:
    - Ruler: inches or centimeters
    - Years of work experience
    - Income: money earned last year

TYPES OF SCALES

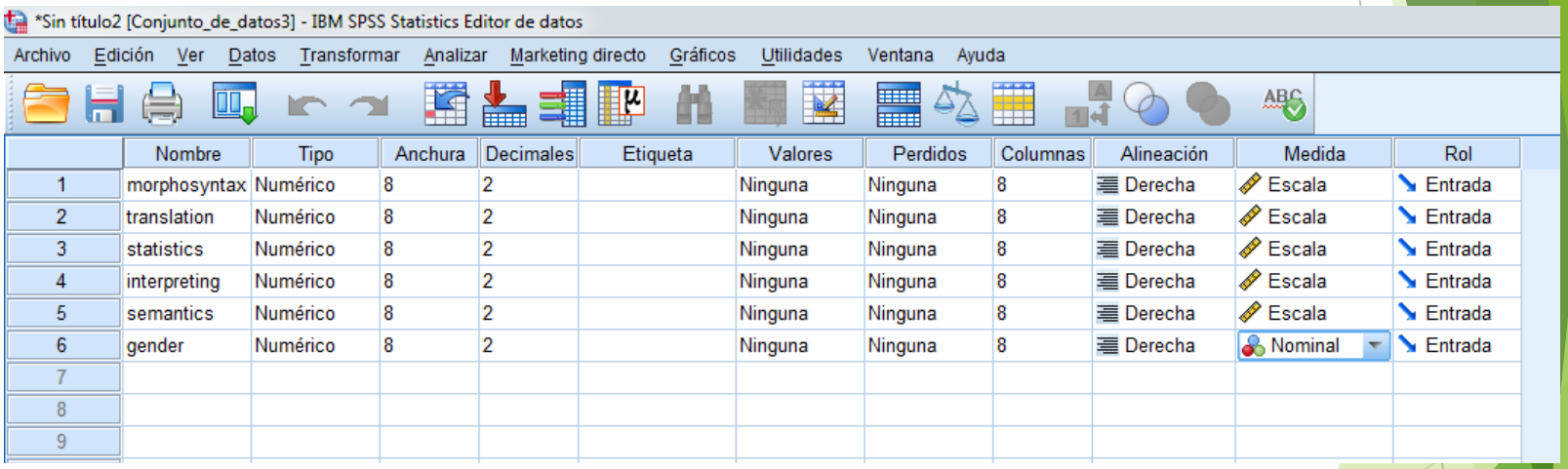
- FURTHER EXAMPLES
  - Satisfaction => ordinal scale
  - Nationality => nominal/qualitative scale
  - Height => rational scale
  - Hair color => nominal/qualitative scale
  - Discomfort => ordinal scale
  - Weight => rational scale

Measures of central tendency

- Interval data
  - roughly normally distributed data (less appropriate for skewed distributions)
    ➔ mean (although mode and median should give the same results)
  - strongly skewed
    ➔ mode, median
- Nominal or categorical data
  ➔ mode

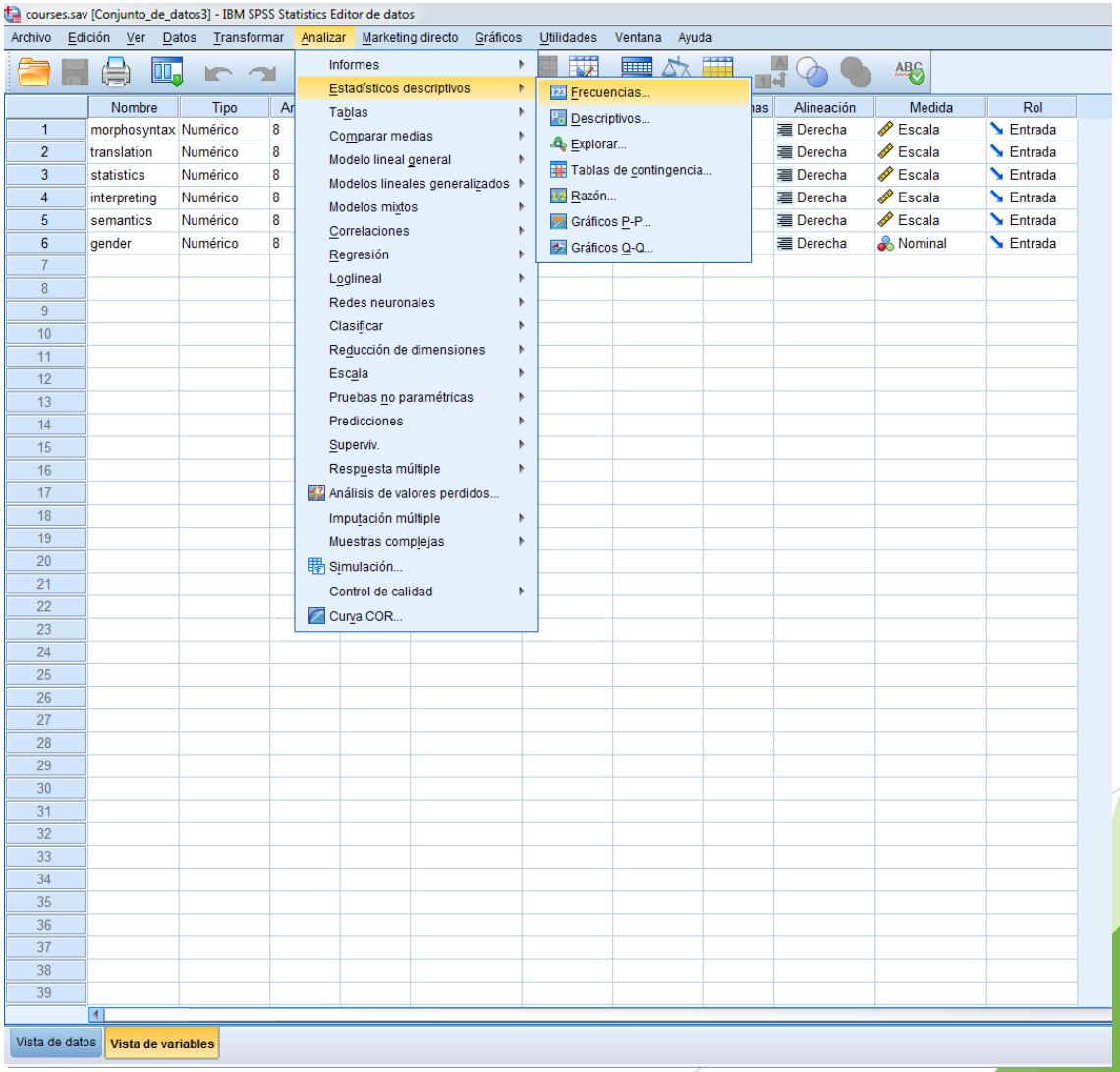
SPSS Statistics

- Software package used for statistical analysis

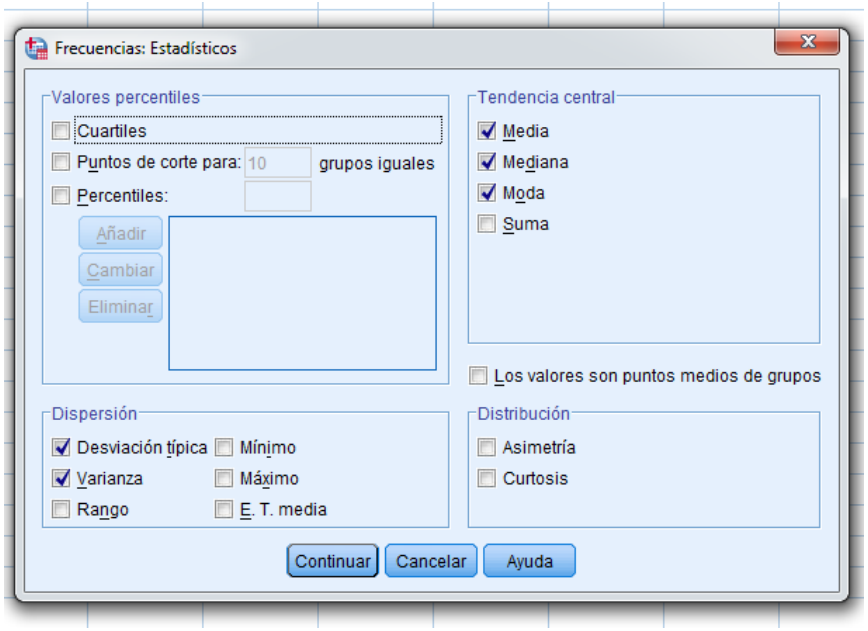


	Nombre	Tipo	Anchura	Decimales	Etiqueta	Valores	Perdidos	Columnas	Alineación	Medida	Rol
1	morphosyntax	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
2	translation	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
3	statistics	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
4	interpreting	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
5	semantics	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Escala	Entrada
6	gender	Numérico	8	2		Ninguna	Ninguna	8	Derecha	Nominal	Entrada
7											
8											
9											

SPSS Statistics



SPSS Statistics

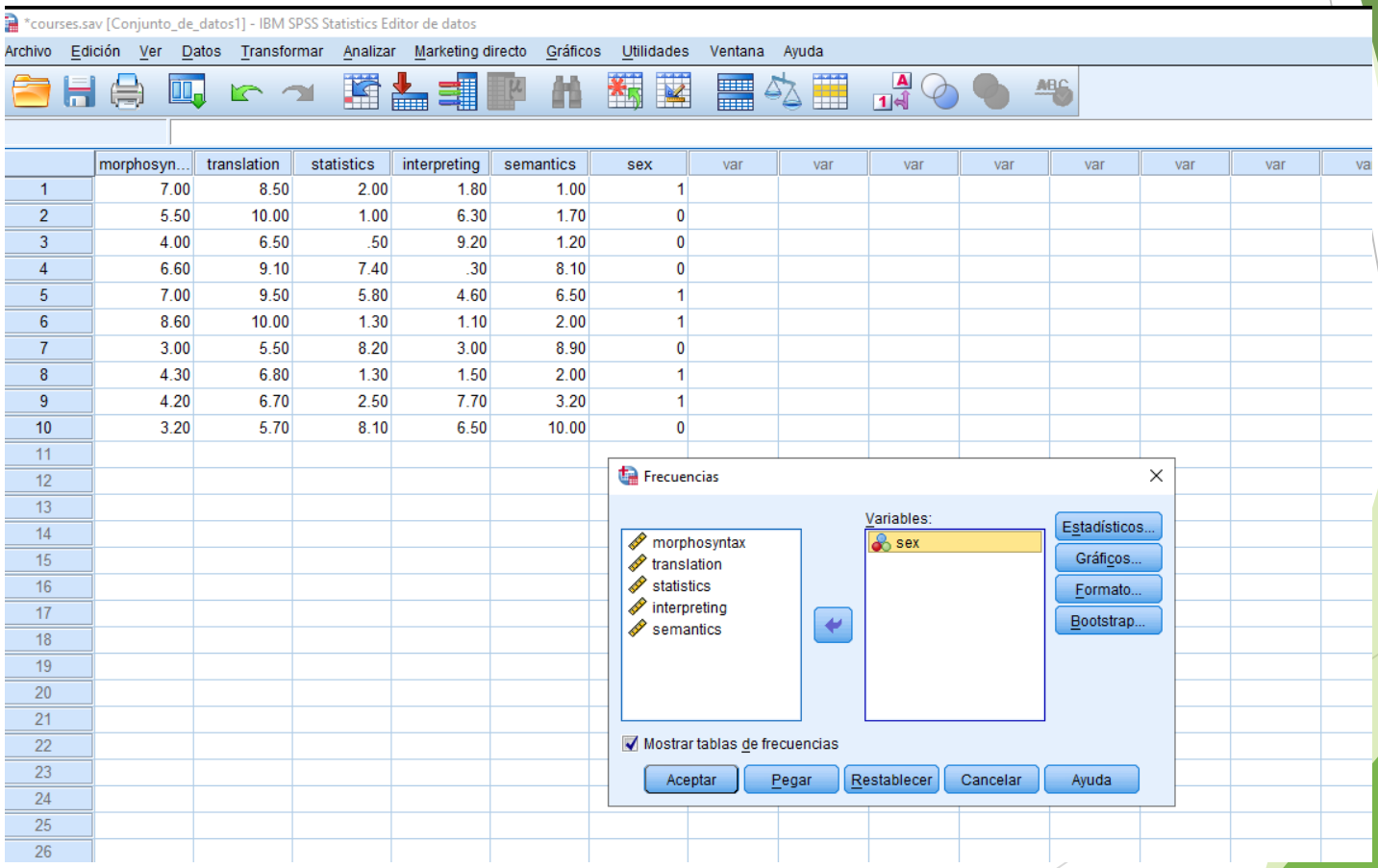


SCALES AND VARIABLES
TYPES OF SCALES

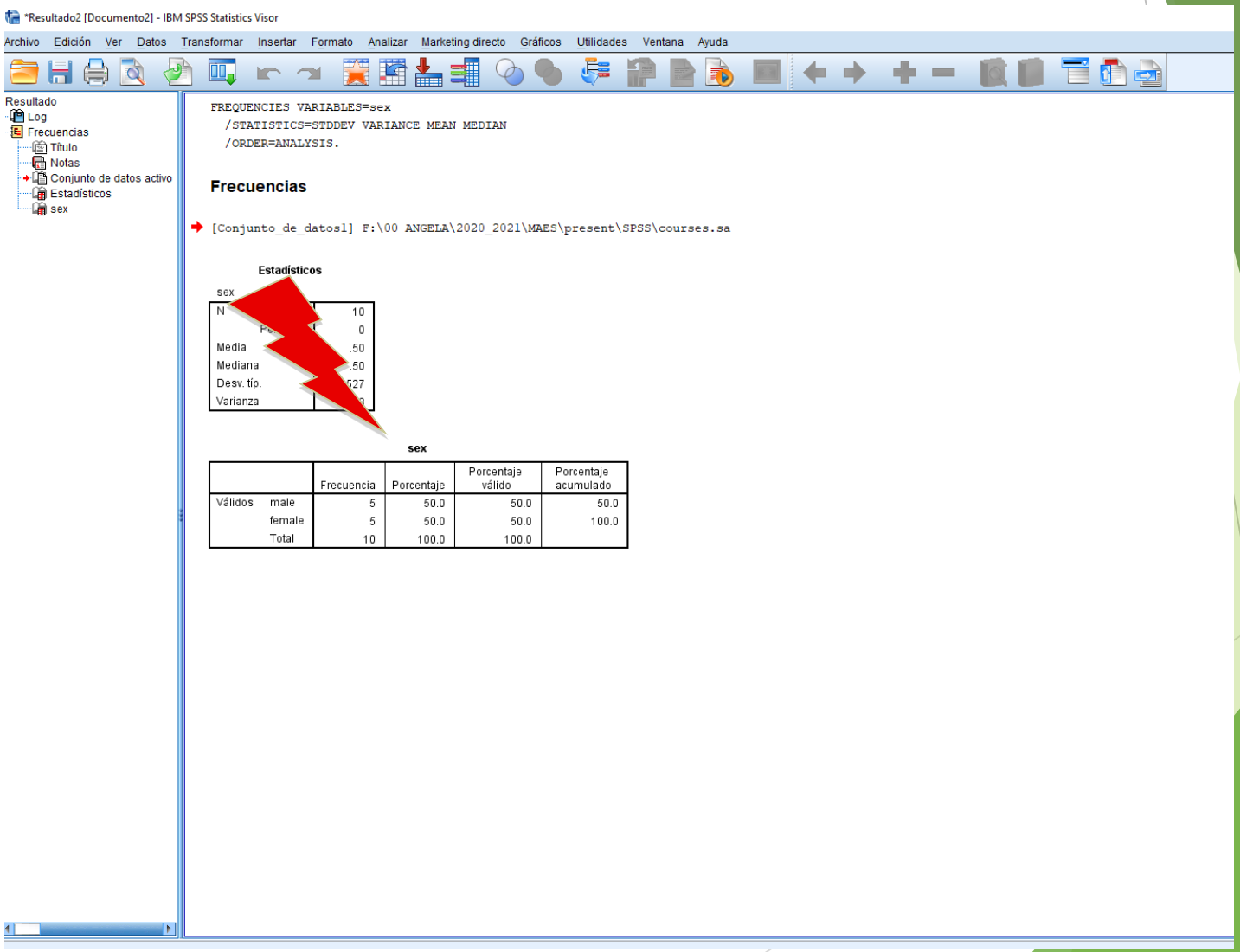
Question:

On which type of data can we calculate a meaningful “central tendency”?

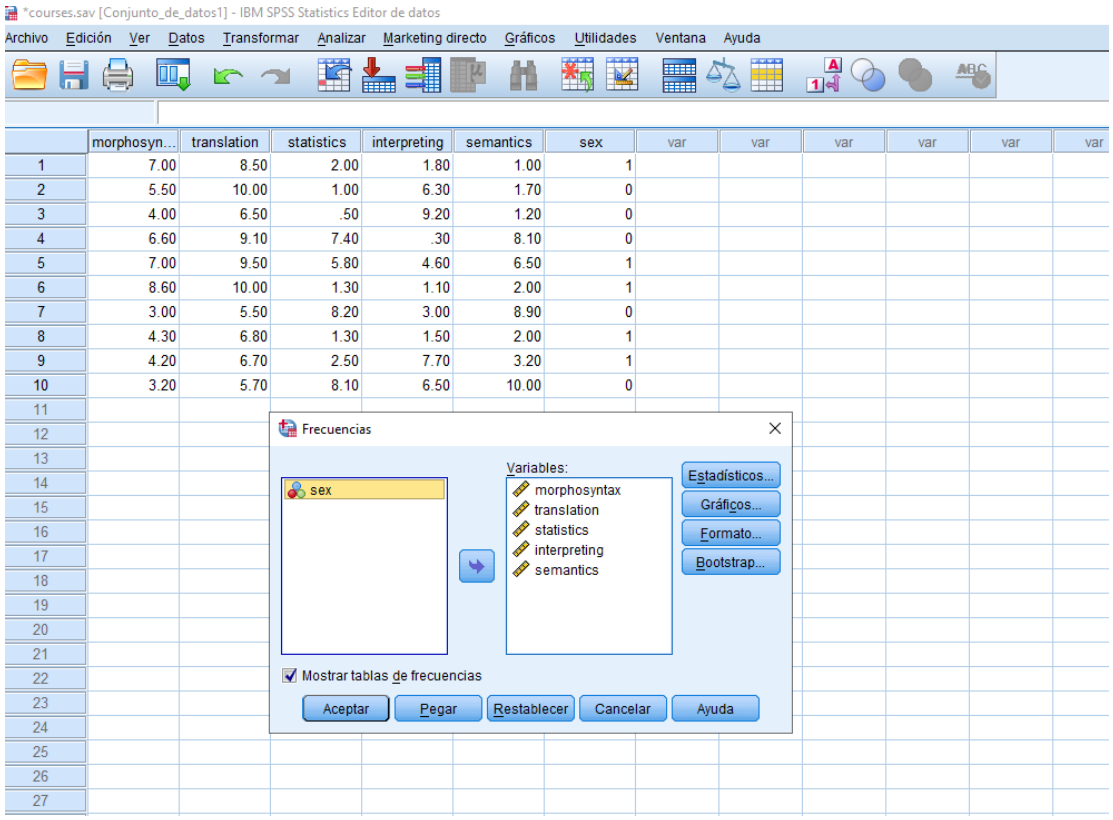
Beware calculations with nominal variables!



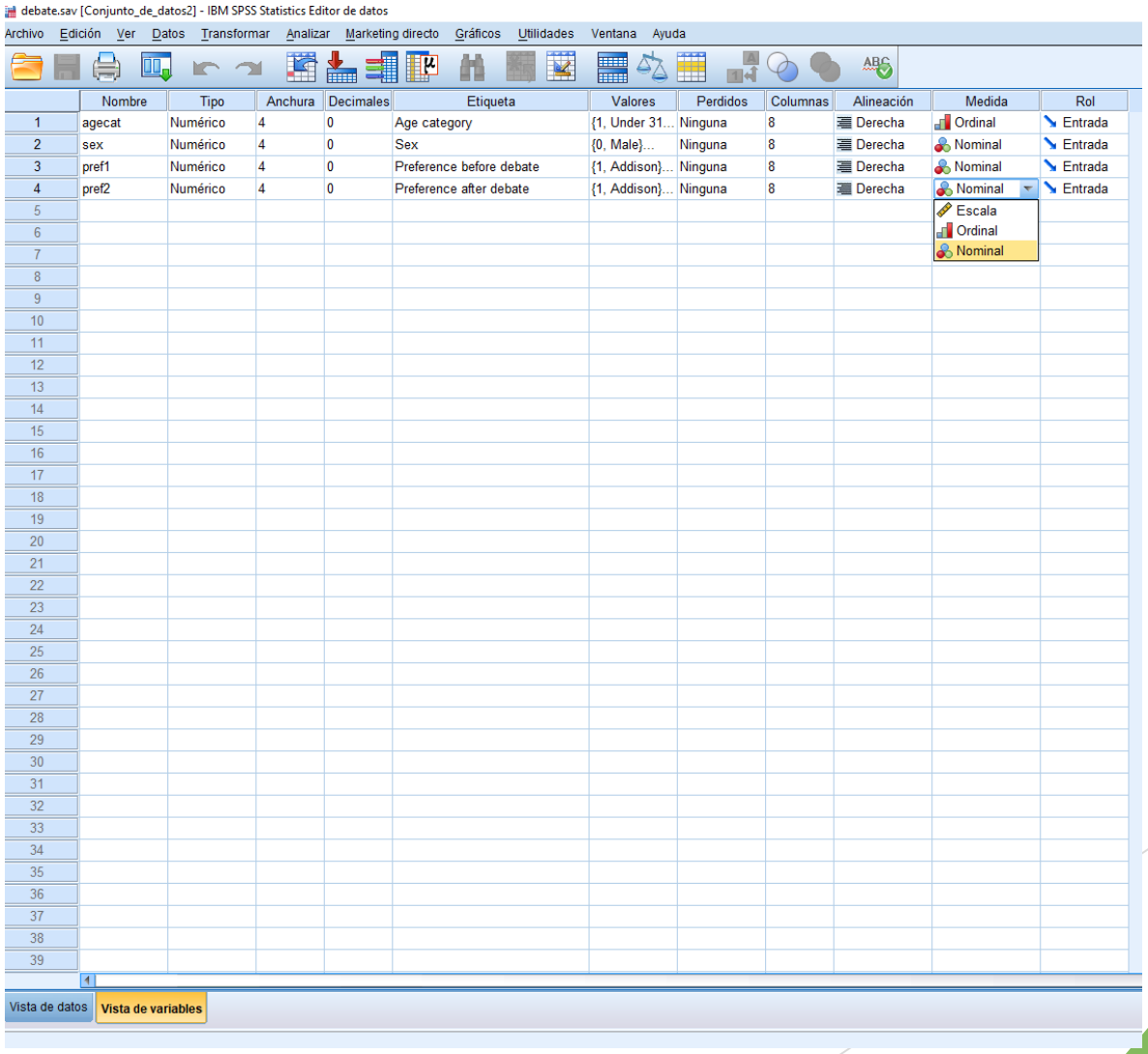
Beware calculations with nominal variables!



Appropriate statistics



SPSS: nominal and ordinal data



SPSS: nominal and ordinal data

Frecuencias

[ConjuntoDatos3] C:\Program Files (x86)\IBM\SPSS\Statistics\23\Samples\English\debate.sav

Estadísticos

Age category

N	Válido	20
	Perdidos	
Media		
Mediana		
Moda		2[a]
Desviación estándar		1,040
Varianza		1,082
Rango		3
Mínimo		1
Máximo		4

a. Existen múltiples modos. Se muestra el valor más pequeño.

Beware calculations with ordinal and nominal variables!

Age category

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	Under 31	3	15,0	15,0	15,0
	31-45	6	30,0	30,0	45,0
	46-60	6	30,0	30,0	75,0
	Over 60	5	25,0	25,0	100,0
	Total	20	100,0	100,0	

SCALES AND VARIABLES
Experimental research

- Often, we are interested in whether human behavior is dependent on certain factors, e.g.
  - Is average sentence length dependent on the dialectal region?
  - Does biological sex affect lexical choice?

SCALES AND VARIABLES
Dependent and independent variables

- Independent variable:
  - Variable(s) manipulated by the experimenter
  - Experimenter determines the values it will assume
  - Independent variables may have a number of different levels
- Dependent variable:
  - Measure of behavior (not manipulated or controlled by experimenter)

Some practice

What are the dependent and independent variables in this research question?

- Does sex influence speech rate in adults?

Some practice

What are the dependent and independent variables in this research question?

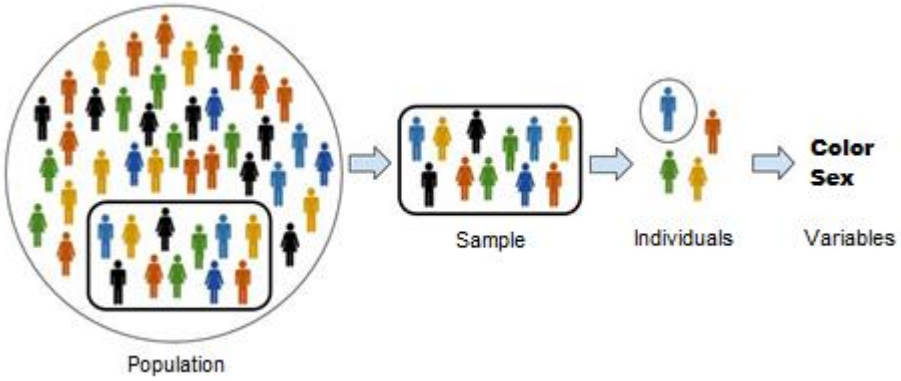
- Does sex influence speech rate in adults?
  - dependent variable: speech rate (interval/rational data)
  - independent variable: sex (2 levels)

Index

- Why using statistics in linguistic research?
- Some basic notions: descriptive statistics
  - Central tendency
    - Mode
    - Median
    - Mean
  - Measures of variability or dispersion
    - Variance
    - Standard deviation
- Scales and variables
- Basic notions of inferential statistics

Inferential statistics

- Inferential statistical analysis infers properties of a population, for example by testing hypotheses
  - It is assumed that the observed data set is sampled from a larger population



Hypothesis testing: 4 steps
(Cantos, 2013:44)

- 1. Formulate the null hypothesis H_0 and the alternative hypothesis H_1
- 2. Identify the test statistics that can be used to assess the truth of the null hypothesis H_0
  - Characteristics of the data, parameterized distribution, etc.
- 3. Compute the *p-value*, which is the probability that a test statistic at least as significant as the one observed would be obtained assuming that the null hypothesis H_0 were true.
  - **The smaller the *p-value*, the stronger the evidence against the null hypothesis.**
- 4. Compare the *p-value* to an acceptable significance value *a*. If $p \leq a$:
  - the observed effect is statistically significant
  - the null hypothesis is ruled out
  - the alternative hypothesis is valid

Hypothesis testing

- Null-hypothesis H_0
  - Generally phrased to negate the possiblity of a relationship between the independent and dependent variables.
  - If the null-hypothesis is true, there is no interaction between dependent and independent variables.
  - Alternative hypothesis contradicts null-hypothesis

State the null hypothesis in the following research question:

- Does gender influence speech rate in adults?
  - dependent variable: speech rate (interval/rational data)
  - independent variable: gender (2 levels)

  - null-hypothesis: there is no difference in speech rate for males and females

Statistical significance

- Significance tests allow us to determine whether or not a finding is…
  - The result of a genuine difference between 2 (or more) items;
  - Just due to chance

- It has been somewhat arbitrarily defined by statisticians as the probability of phenomena occurring due to chance less than 5% ($p < .05$)

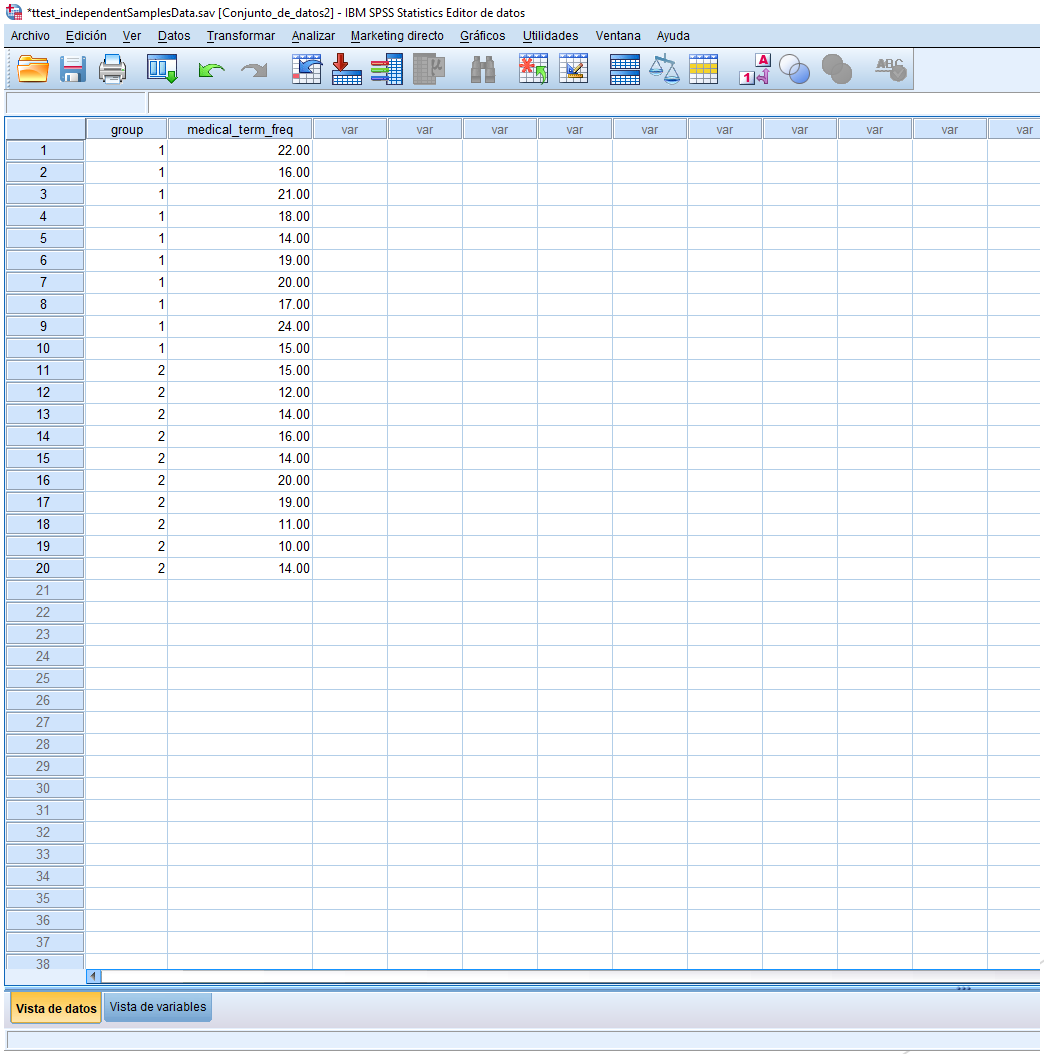
Hands-on Quantitative Methods Research: Practical examples

Ángela Almela, PhD
Universidad de Murcia

Introduction to parametric versus non-parametric statistics

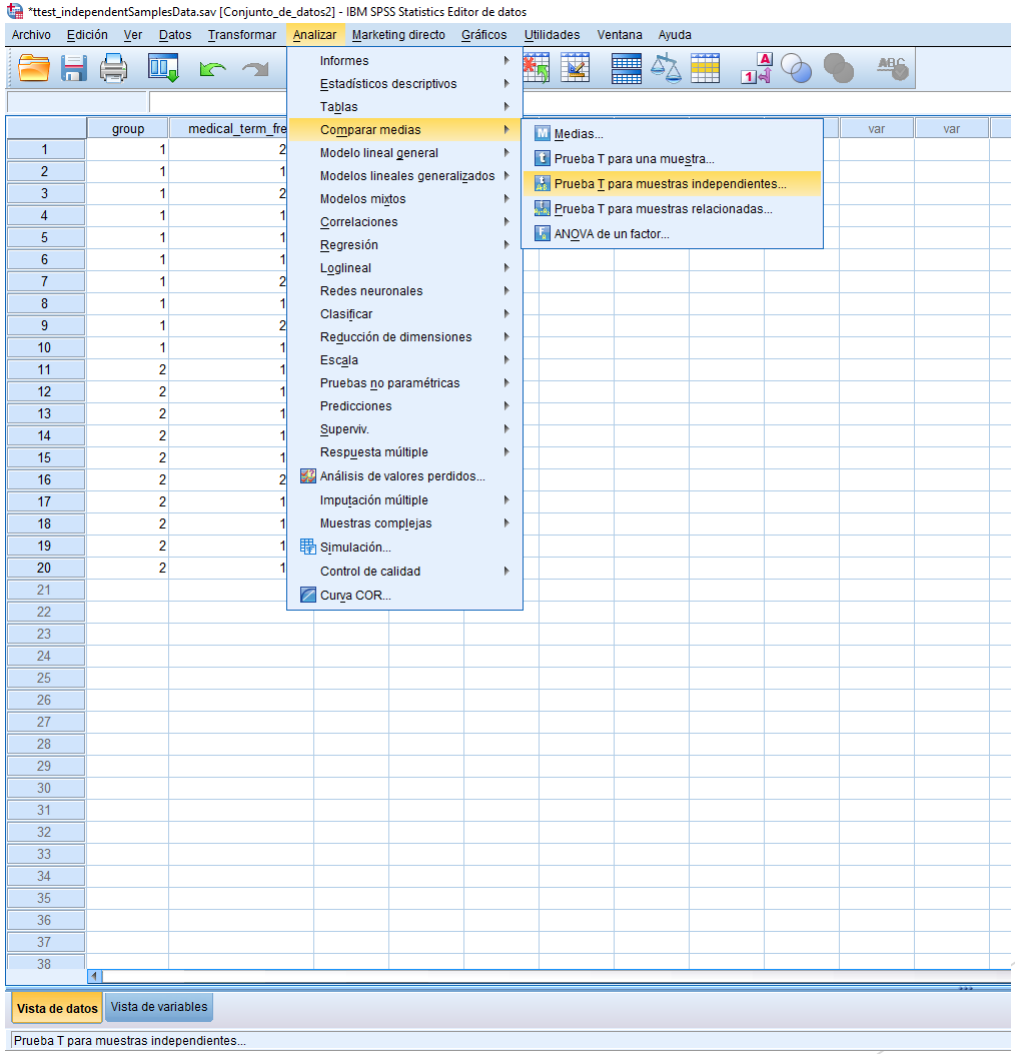
- Tests of differences between groups (independent samples). Two groups
  - *t-test for independent samples* (§3.2.1, Cantos 2013)
- Tests of differences between variables (dependent samples)
  - *t-test for paired samples* (§3.2.2, Cantos 2013)

IBM SPSS Statistics – Tests of differences between groups (independent samples)

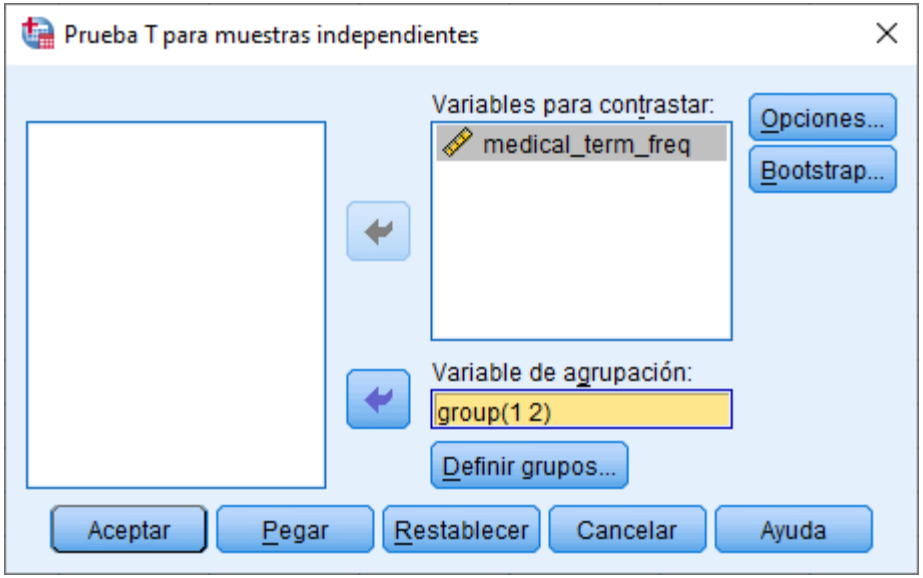


	group	medical_term_freq	var	var	var	var	var	var	var	var	var	var
1	1	22.00										
2	1	16.00										
3	1	21.00										
4	1	18.00										
5	1	14.00										
6	1	19.00										
7	1	20.00										
8	1	17.00										
9	1	24.00										
10	1	15.00										
11	2	15.00										
12	2	12.00										
13	2	14.00										
14	2	16.00										
15	2	14.00										
16	2	20.00										
17	2	19.00										
18	2	11.00										
19	2	10.00										
20	2	14.00										
21												
22												
23												
24												
25												
26												
27												
28												
29												
30												
31												
32												
33												
34												
35												
36												
37												
38												

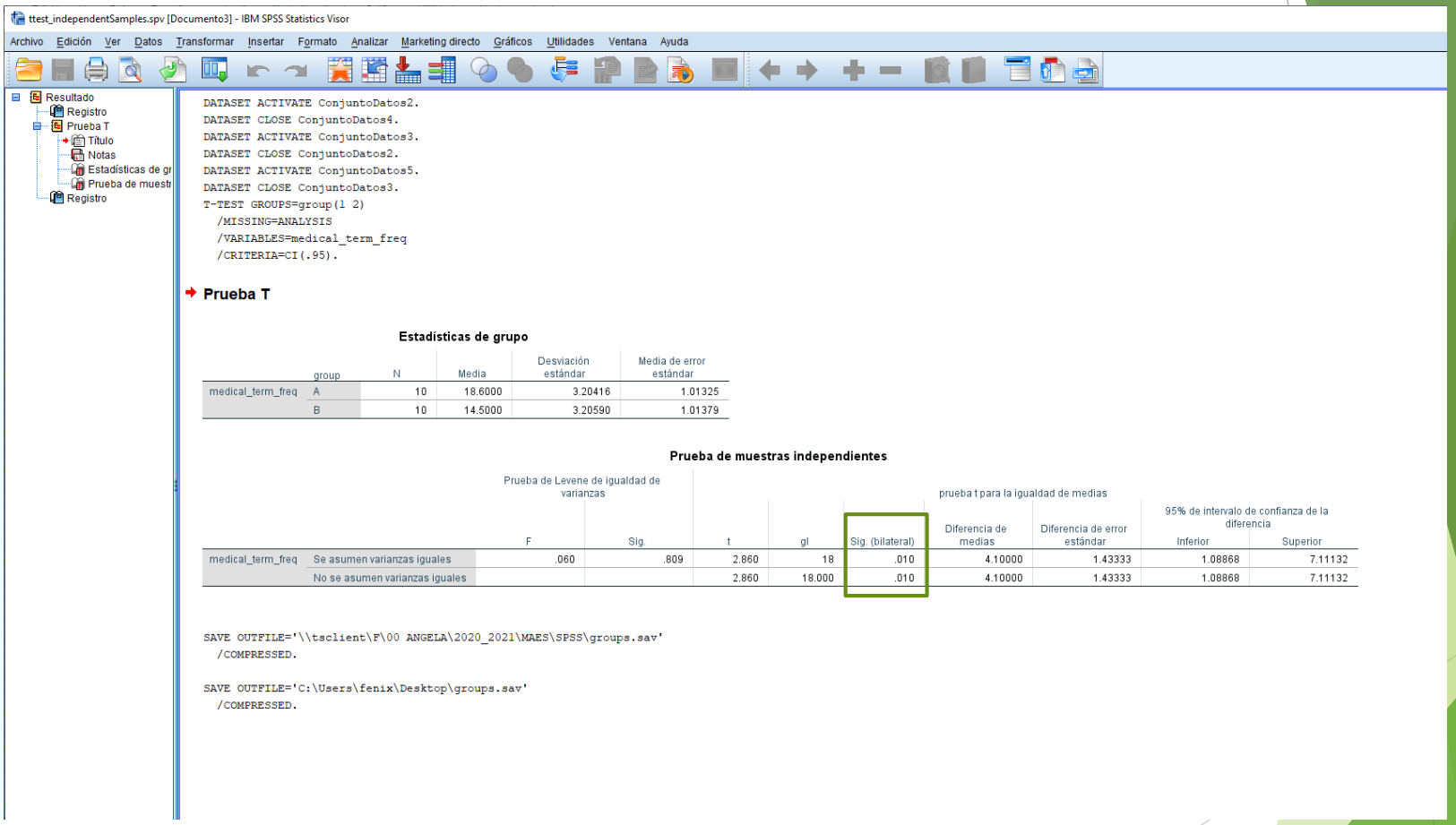
IBM SPSS Statistics – Tests of differences between groups (independent samples)



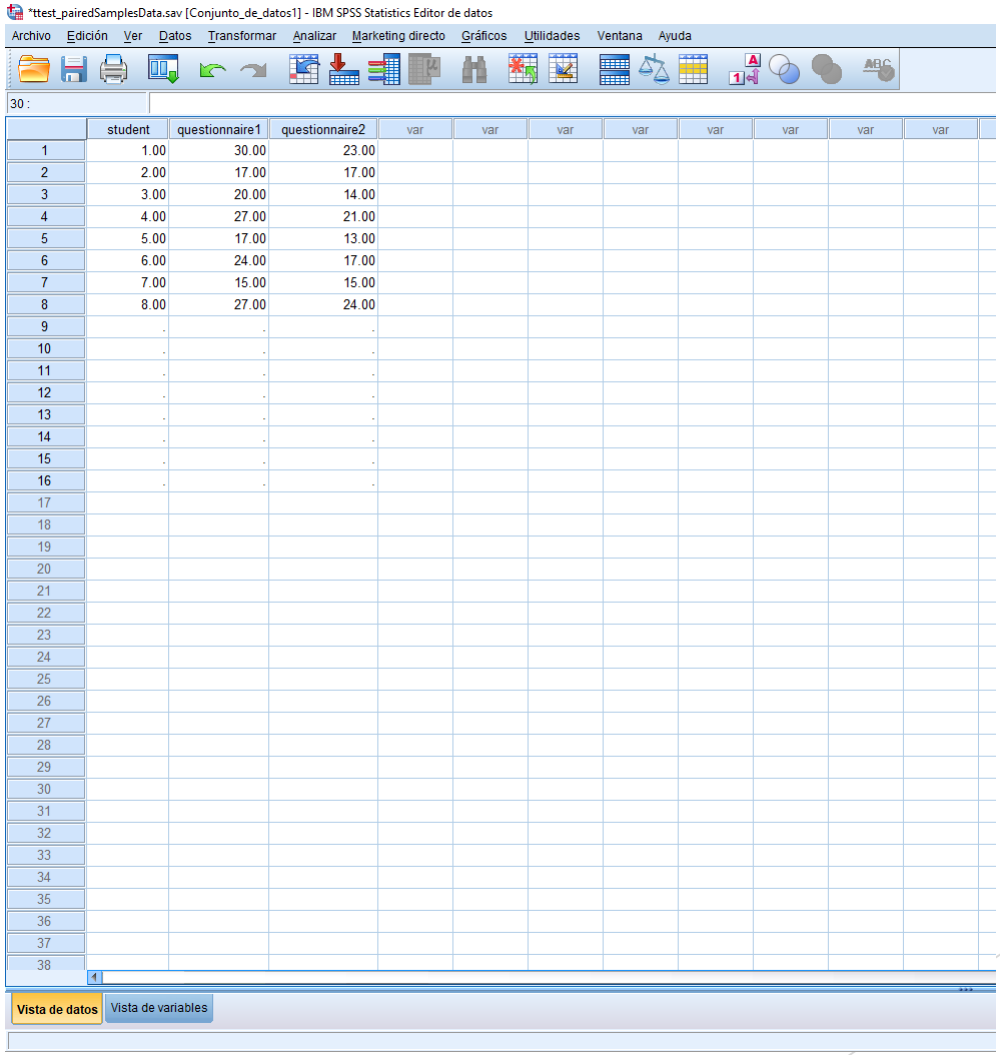
IBM SPSS Statistics – Tests of differences between groups (independent samples)



IBM SPSS Statistics – Tests of differences between groups (independent samples)

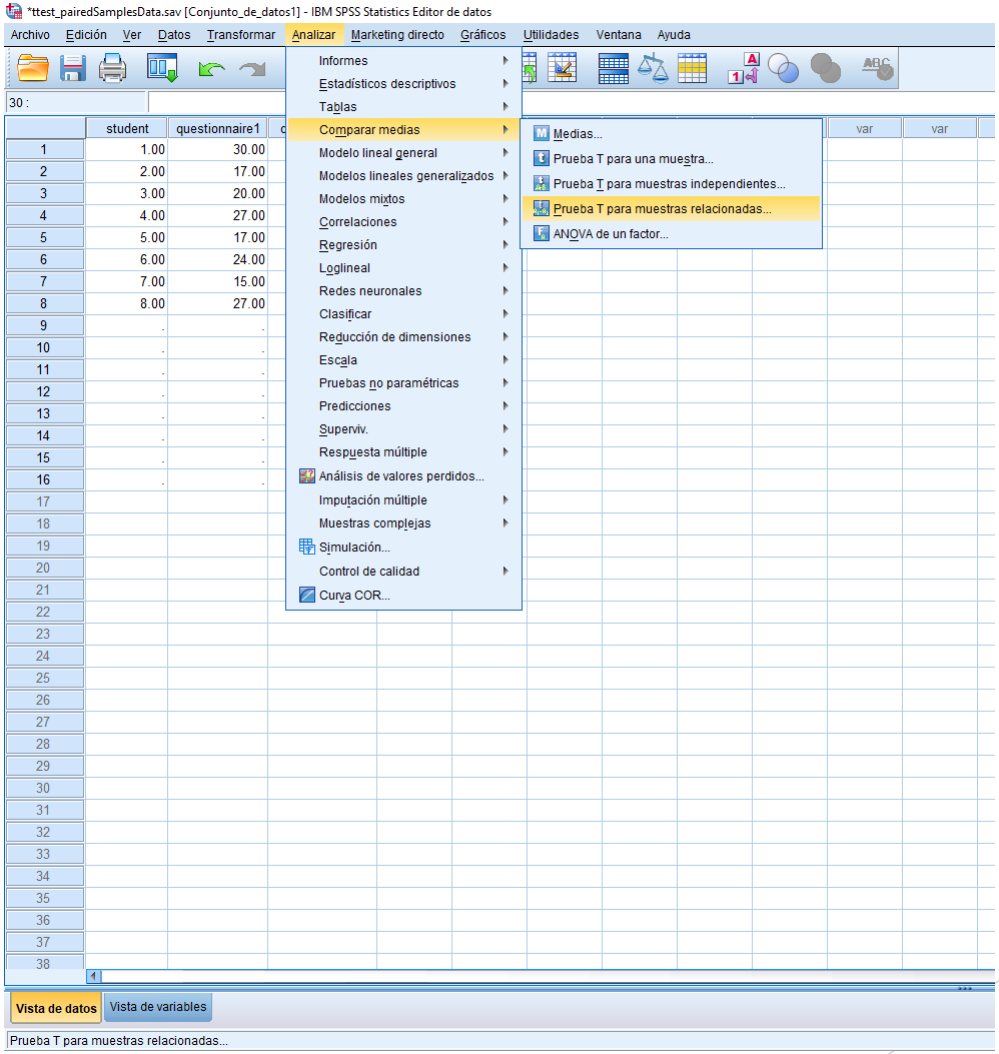


IBM SPSS Statistics – Tests of differences between variables (dependent samples)

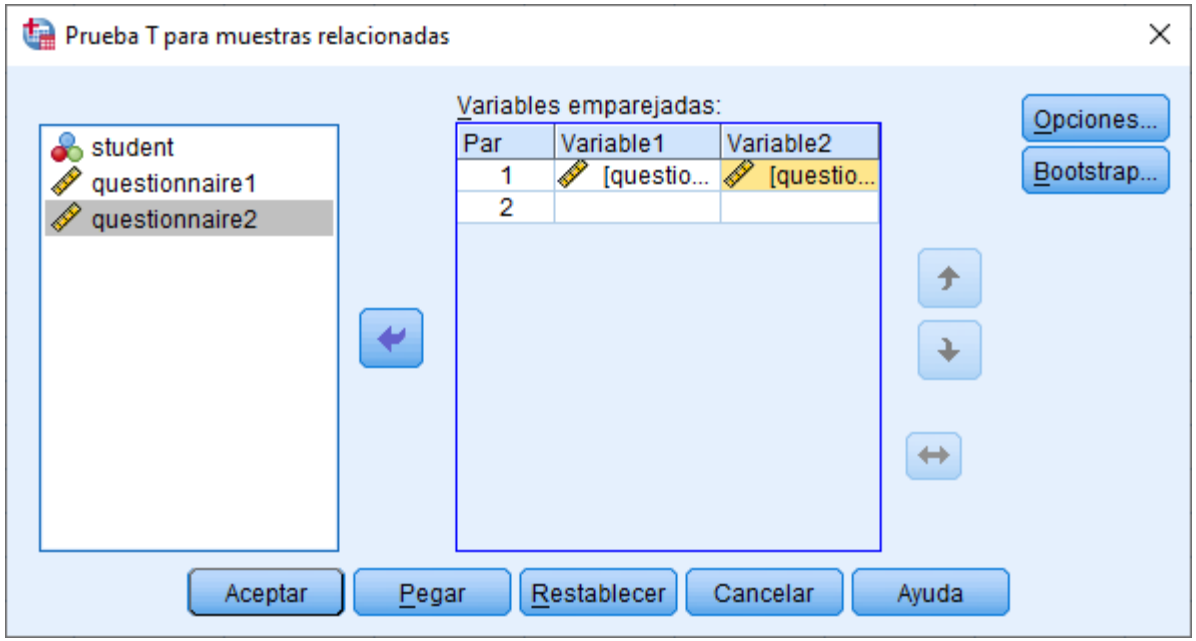


	student	questionnaire1	questionnaire2	var	var	var	var	var	var	var	var
1	1.00	30.00	23.00								
2	2.00	17.00	17.00								
3	3.00	20.00	14.00								
4	4.00	27.00	21.00								
5	5.00	17.00	13.00								
6	6.00	24.00	17.00								
7	7.00	15.00	15.00								
8	8.00	27.00	24.00								
9	.	.	.								
10	.	.	.								
11	.	.	.								
12	.	.	.								
13	.	.	.								
14	.	.	.								
15	.	.	.								
16	.	.	.								
17											
18											
19											
20											
21											
22											
23											
24											
25											
26											
27											
28											
29											
30											
31											
32											
33											
34											
35											
36											
37											
38											

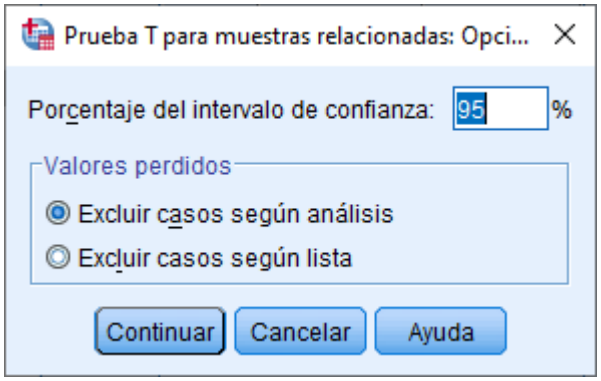
IBM SPSS Statistics – Tests of differences between variables (dependent samples)



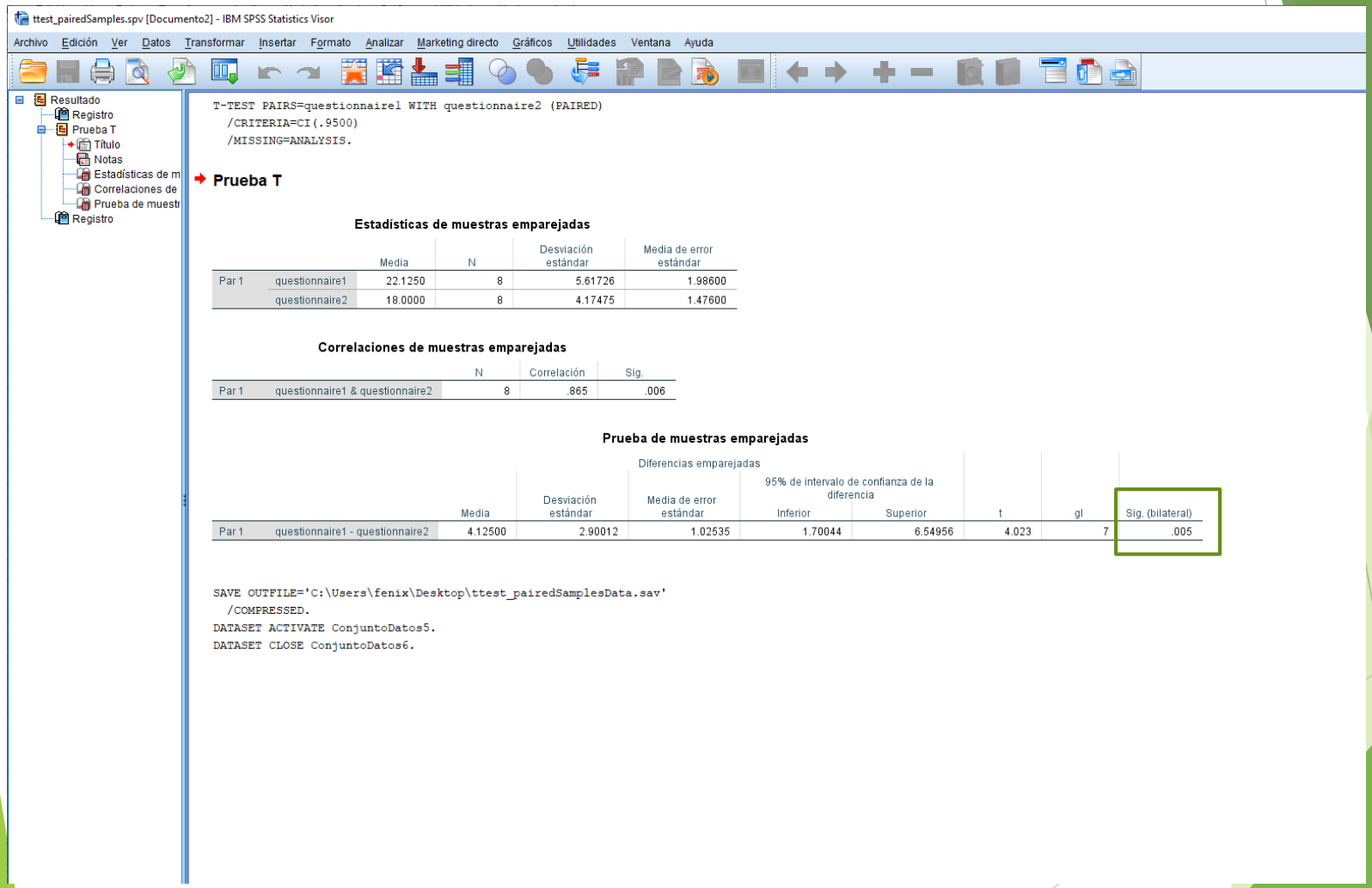
IBM SPSS Statistics – Tests of differences between variables (dependent samples)



IBM SPSS Statistics – Tests of differences between variables (dependent samples)



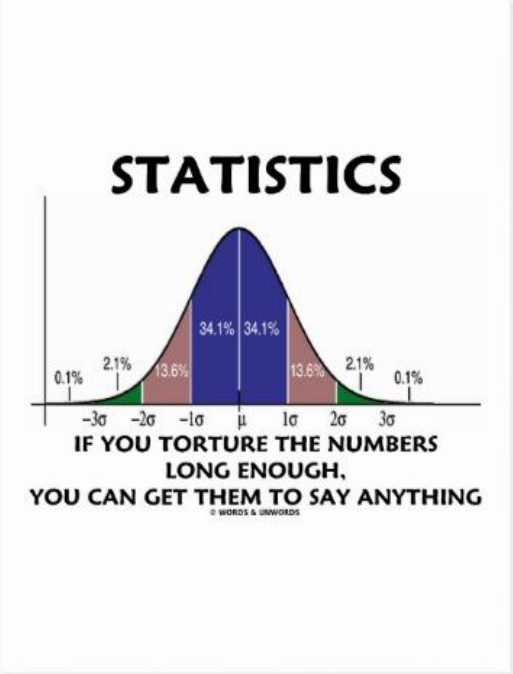
IBM SPSS Statistics – Tests of differences between variables (dependent samples)



Questions to be addressed when evaluating Discussion/Conclusion

1. Do the findings logically answer the research questions or support the research hypothesis?
2. Does the nature of the study remain consistent from beginning to end?
3. Are the findings generalized to the correct population or situations?
4. Are the conclusions consistent with the type of research design used?
5. Are the findings and conclusions related to theory or previous research?
6. Are the limitations of the study made clear?
7. Is there consistency between the findings and the applications?

In a nutshell…



Bibliography and materials used

- Braun, B. (2004). Class in support of the Lecture "Statistical Methods for the Social Sciences." Retrieved from http://www.phon.ox.ac.uk/files/people/braun/teaching.html
- Cantos, P. (2013). *Statistical Methods in Language and Linguistic Research*. Sheffield, UK: Equinox Publishing.
- Gries, S.Th. (2021). *Statistics for linguistics with R: A practical introduction* (3rd edition). Berlin: De Gruyter Mouton.
- Intuitor (2001). *Type I and Type II Errors - Making Mistakes in the Justice System*. Retrieved from http://intuitor.com/statistics/T1T2Errors.html
- Levshina, N. (2015). *How to do linguistics with R: Data exploration and statistical analysis*. John Benjamins Publishing Company.
- Patwari, R. (2013, September 14). *Hypothesis Testing 01: Going To Court* [video file]. Retrieved from https://www.youtube.com/watch?v=_Z5gPXoRkic

Sources and recommended websites about statistics

- https://www.onlinemathlearning.com/median.html
- https://www.mymarketresearchmethods.com/types-of-data-nominal-ordinal-interval-ratio/
- https://www.khanacademy.org/math/ap-statistics/summarizing-quantitative-data-ap/measuring-center-quantitative/e/calculating-the-median

Image sources

- http://ww1.cazoomaths.com/
- https://www.wallstreetmojo.com/variance-vs-standard-deviation/
- http://www.amathsdictionaryforkids.com/qr/f/frequencyDistributionTable.html
- http://www.picturequotes.com/statistics-can-be-made-to-prove-anything-even-the-truth-quote-47723
- https://www.zazzle.com/torture+the+numbers+gifts
- https://www.editoresweb.es/sinonimo-de-importante/
- http://www.regentsprep.org/regents/math/algtrig/ats2/normallesson.htm
- https://www.onlinemathlearning.com/median.html
- https://media.giphy.com/media/10bv4HhibS9nZC/giphy.gif
- https://www.datalaria.com/en/post/fundamentos/2018-10-07-estadistica-descriptiva/

Thanks for your attention!

angelalm@um.es