Research Methods for Political Science PO3110 (TCD)

HT: Tutorial 2 - Week 3

Letícia Meniconi Barbabela

University College Dublin,
https://github.com/letmeni/research-methods

4-5 February 2020

Last week: Hypothesis testing

- Prediction from theory, eg.: *difference with respect to a set value, relationship between variables*;
- Fit a model to the data and evaluate the probability of the results shown by the model given the assumption that no effect exists (null hypothesis);

$$outcome_i = bX_i + error_i$$

  - Prediction: difference (one variable) or relationship (two variables)?
  - Level of measurement of variable(s);
  - Choose levels of significance (eg.: 95%);
  - Test statistic $= \frac{effect}{error}$
  - p-value: probability of getting such a test statistic score under null hypothesis.

Some statistical tests

- **One sample t-test:** Difference between the mean of a variable and a constant value;
- χ^2 **and Cramer's V:** Association between two categorical variables;
- **Correlation:** Association between two continuous variables.

One Sample T-test

Compare the mean of a continuous variable to a specified constant value[1], e.g.:
Do students from this class have grades higher than 75%?

- H_0: there isn't a difference between the observed value and the reference one;
- H_1: there is a difference between the observed value and the reference one;
- Evaluate whether it is a one-tailed t-test (directional: in our example higher or lower grades) or a two-tailed one (non-directional: in our example different grades);

$$t = \frac{\text{observed value - expected value under } H_0}{\text{standard error}} = \frac{\bar{x} - m_0}{s/\sqrt{n}}$$

[1] the comparison value (m_0) could be the population mean (μ)

χ^2

Independence between two categorical variables, e.g.: *Do Tuesday students wear black sweaters more often than Wednesday ones?*

- H_0: x is independent upon y
- H_1: x is dependant upon y
- We need to know 2 things: the χ^2 score and the degrees of freedom (df):

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

  - f_o = observed frequencies
  - f_e = expected frequencies (assuming independence) = $\frac{\text{row margin} * \text{column margin}}{\text{total}}$

$$df = (\text{rows} - 1) * (\text{columns} - 1)$$

- Tell us if we can reject the null hypothesis about independence, but nothing about the strengh of the relationship (see Cramer's V)

Cramer's V

- Strengh of relationship between two categorical variables:

$$V = \sqrt{\frac{\chi^2}{N*\text{k-1}}}$$

- k = rows (r) or collumns (c), whichever is smaller;
- If N is large, you are likely to find a significant relationship (but it might be a weak one).

Other measures:

Go back to MT7 slides for measures of association λ and γ

Variance, Co-variance and Correlation

Relationship between two numeric variables, e.g.: *Is studying more hours associated to having higher grades?* :

- Var(x) = Cov(x,x)
  - Variance: $\sigma^2 = \frac{\sum(x-\bar{x})^2}{n-1} = \frac{\sum(x-\bar{x})(x-\bar{x})}{n-1}$
  - Co-variance: $\sigma_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$
  - $\sigma_x\sigma_y \leq \sigma_{xy} \leq \sigma_x\sigma_y$
- Correlation: Co-variance standardized
  - $r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$
  - $-1 \leq r \leq 1$

Statistical tests on SPSS

1. Let's go back to analysing the conflict dataset on SPSS;
   - `https://tinyurl.com/method-conflict`
2. Define a research question and a hypothesis;
3. Describe the variable(s) you are interested at using plots and/or tables;
4. Identify and perform a suitable statistical test;
5. Most important part: Interpret the results.

Simple linear regression (Field 2013)

- Linear regression: equation of a straight line;
- "Simple": only one independent variable;
- This model differs from that of a correlation only in that is uses an unstandardized measure of the relationship (β) and consequently we need to include a parameter that tell us the value of the outcome when the predictor is zero. This parameter is β_0;
- This is a useful tool, but there are some assumptions (we will go over them in a few tutorials)...

Correlation and regression

BASIS FOR COMPARISON	CORRELATION	REGRESSION
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

https://www.datasciencecentral.com/

Regression equation

- General idea: choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- $SSR = \sum(Y_i + \hat{\beta}_0 + \hat{\beta}_1 \times X_i)^2$
- R^2 the root-mean squared error, calculated as $\sqrt{\frac{1}{n} \times SSR}$
- R square indicates how much (ratio) of the variance in the dependent variable can be explained by our regression model

Simple regression on SPSS

Since we already have the conflict dataset open ...

1. Go back to the correlation we looked at before;
2. Let's make a simple regression out of it;
3. What are the important things to look at (at this point)?
4. How do we interpret the output?

Simple regression on SPSS

Since we already have the conflict dataset open ...

1. Go back to the correlation we looked at before;
2. Let's make a simple regression out of it;
3. What are the important things to look at (at this point)?
4. How do we interpret the output?

Simple regression on SPSS

Since we already have the conflict dataset open ...

1. Go back to the correlation we looked at before;
2. Let's make a simple regression out of it;
3. What are the important things to look at (at this point)?
4. How do we interpret the output?

Some definitions... (Field 2013)

- "R^2 tells us how much variance is explained by the model compared to how much variance there is to explain in the first place. It is the proportion of variance in the outcome variable that is shared by the predictor variable."
- "**F** tells us how much variability the model can explain relative to how much it can't explain (i.e., it's the ratio of how good the model is compared to how bad it is)."

Anova box

- includes tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'
- F-ratio represents the ratio of the improvement in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model
- F-ratio is calculated by dividing the average improvement in prediction by the model (MS_M) by the average difference between the model and the observed data (MS_R).
- If the improvement due to fitting the regression model is much greater than the inaccuracy within the model, then the value of F will be greater than 1, and SPSS calculates the exact probability of obtaining the value of F by chance

How to present your homework assignments

- Firstly, make sure you have working a syntax file.
- Secondly, run it!
- Export the output (no screen shots);
- Write a comment interpreting the outcome if required.

Other minor things:

- Use mathematical notation: x bar is $\bar{x}$, mu is μ, var(x) is σ^2 etc;
- Make sure the document you are working on is clear and legible;
- Go back to tutorial exercises.

References

- Field, A (2013) *Discovering Statistics Using SPSS*. 4th edition. London:Sage
- HT 2019 Slides at `http://andrsalvi.github.io/research-methods`