Research Methods for Political Science PO3110 (TCD)

HT: Tutorial 1 - Week 2

Letícia Meniconi Barbabela

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

28-29 January 2020

Tutorials

- Participation is mandatory;
- Come prepared: reading assigned materials, attending lectures, doing homework;
- Using SPSS (Download it here);
- Going over homework, problem sets and topics from lectures (not a substitute!);

Assessment

1. 60% ⇒ Exam;
2. 20% ⇒ 4 homework exercises.
   - Submit online via Blackboard (Turnitin) on the Monday evening preceding the tutorial session;
3. 16% ⇒ Research project paper:
   - Deadline: Monday, 20/04 @ 11:59pm
   - Check assigned groups on Blackboard;
4. 4% ⇒ Tutorial participation, including presentation sessions (Weeks 11&12):
   - *Two unexcused absences in tutorials will be tolerated. Beyond that, the student will receive a zero for participation.*

Late Submission Policy

- Talk to me in advance and send evidence of serious circumstances;
- 5 points per day will be taken off your mark on assignments submitted late without a valid and previous excuse (capped at 30 points for the paper);
- Homework exercices received after noon on Tuesdays automatically receives a zero.

Assignments' Submission

- Blackboard/Turnitin (not by email!);
- LaTeX, Word/Open Office and submitted as **PDFs**;
- Statistical Software: SPSS. You can use alternatives such as R or STATA if you want, but not Excel!
- Please do include the syntax (code) from whichever software you are using;
- When including tables use the "export" function from SPSS saving figures in high resolution;
- Don't submit screen-shots.

Important Dates

Homework:

- Week 4: HW 1 (Monday, 10 February @ 11:59pm)
- Week 6: HW 2 (Monday, 24 February @ 11:59pm)
- Week 9: HW 3 (Monday, 23 March @ 11:59pm)
- Week 11: HW 4 (Monday, 6 April @ 11:59pm)

Research project:

- Weeks 11 & 12: Group presentations;
- Monday, 20 April @ 11:59pm: Paper submission.

Support

- Slides from HT & MT 2019 :
  `http://andrsalvi.github.io/research-methods`
- Slides from **our tutorials** (HT 2020):
  `https://github.com/letmeni/research-methods`
- Questions:
  1. preferably in class;
  2. Office hours: by appointment, preferably on Tuesdays 11-12;
  3. leticia.barbabela@ucdconnect.ie

Today's tutorial

- Describe/summarize a **sample**: measures of central tendency and dispersion;
- Infer something about **population**;
- **Test** a hypothesis.

Using SPSS to describe distributions of variables in our sample

1. Open SPSS;
2. Download the dataset from James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War," American Political Science Review 97, 1 (March 2003): 75-90:
   - https://tinyurl.com/method-conflict
3. Calculate mode, median, mean and standard deviation for "population":
   - What do we conclude?
4. Do the same for "country region":
   - What is the problem?

Using SPSS to describe distributions of variables in our sample

1. Open SPSS;
2. Download the dataset from James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War," American Political Science Review 97, 1 (March 2003): 75-90:
   - `https://tinyurl.com/method-conflict`
3. Calculate mode, median, mean and standard deviation for "population":
   - What do we conclude?
4. Do the same for "country region":
   - What is the problem?

Using SPSS to describe distributions of variables in our sample

1. Open SPSS;
2. Download the dataset from James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War," American Political Science Review 97, 1 (March 2003): 75-90:
   - `https://tinyurl.com/method-conflict`
3. Calculate mode, median, mean and standard deviation for "population":
   - What do we conclude?
4. Do the same for "country region":
   - What is the problem?

Using SPSS to describe distributions of variables in our sample

1. Open SPSS;
2. Download the dataset from James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War," American Political Science Review 97, 1 (March 2003): 75-90:
   - `https://tinyurl.com/method-conflict`
3. Calculate mode, median, mean and standard deviation for "population":
   - What do we conclude?
4. Do the same for "country region":
   - What is the problem?

Using SPSS to describe distributions of variables in our sample

1. Open SPSS;
2. Download the dataset from James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War," American Political Science Review 97, 1 (March 2003): 75-90:
   - `https://tinyurl.com/method-conflict`
3. Calculate mode, median, mean and standard deviation for "population":
   - What do we conclude?
4. Do the same for "country region":
   - What is the problem?

Using SPSS to describe distributions of variables in our sample

1. Open SPSS;
2. Download the dataset from James D. Fearon and David D. Laitin, "Ethnicity, Insurgency, and Civil War," American Political Science Review 97, 1 (March 2003): 75-90:
   - `https://tinyurl.com/method-conflict`
3. Calculate mode, median, mean and standard deviation for "population":
   - What do we conclude?
4. Do the same for "country region":
   - What is the problem?

Measures of Central Tendency

Central Tendency:

- Gives us a sense of the where to locate the "centre" of the distribution.

Measures:

1. Mode
2. Median
3. Mean

Practical Calculations of Central Tendency

- **Mode:**
  - The score that occurs most frequently in the data set;
  - The tallest bar in a frequency distribution.
- **Median:**
  - Rank scores according to magnitude;
  - Choose the middle one;
  - Odd: $\frac{n+1}{2}$
  - Even: average between the value at position $\frac{n}{2}$ and $\frac{n+1}{2}$
- **Mean:**
  - Average;
  - $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - Influenced by outliers, while mode and median are not.

Measures of Dispersion

- **Range**: Difference between largest and smallest observation;
- **Deviance/Spread**:
  - **Total dispersion**: "Sum of Squared Errors (SS)"

    $$\sum(x-\bar{x})^2$$

  - **Average dispersion**: "Variance"

    $$\text{Var(x)} = \sigma^2 = \frac{SS}{n-1} = \frac{\sum(x-\bar{x})^2}{n-1}$$

  - **Average dispersion squared**: "Standard deviation"

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

The **sample** variance is denoted by s^2 and the **sample** standard deviation by s.

The goal is inference

More than being able to describe/summarize the **sample** (with measures of central tendency and dispersion), we want to learn something (value, relationship ...) about the **population**, for instance ...

Using known distributions

- In statistics we have some known distributions: t-distribution, chi-square, F-distribution etc;
  - Probability density functions;
- **Normalizing:** Transform our data into a distribution we know, e.g.:

$$z = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard deviation of the estimate}}$$

- we can use this z-score to assess the probability of observing a value this extreme by chance;
- this is more than roughly guessing the probability simply by looking at the distribution in our sample..

Parameter

- A numerical quantity that characterizes a given population;
- As an example we will be looking at the mean;
- The mean is a summary: a hypothetical value that doesn't have to be observed in the data;
- We use the **sample** mean ($\bar{x}$) to estimate the **population** mean (μ), that is, a parameter;
- The sample we have is one of many possible ones (in a distribution of samples, that also can be describe by measures of central tendency and spread);
- Each different sample will have a different mean.

Central Limit Theory

- As samples get large (> 30), the sampling distribution has a **"normal distribution"** with a mean equal to the population mean;
- The standard deviation of the sample means is also known as the "standard error of the mean" or **standard error**:

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

Confidence Intervals

- *Limits constructed such that for a certain percentage of samples (eg.: 95%) the true value of the population parameter will fall within these limits;*
- Confidence Interval (in this case the population parameter is the mean):

  $\text{CI} = \bar{x} \pm z * \frac{s}{\sqrt{n}}$

  The Z score for 95% confidence is 1.96
- It applies to other parameters as well, we are looking at the mean just as an example...
- Let's do it on SPSS with a different dataset...
- Calculate the confidence interval for "age";
- What does this result tell us?

Confidence Intervals

- *Limits constructed such that for a certain percentage of samples (eg.: 95%) the true value of the population parameter will fall within these limits;*

- Confidence Interval (in this case the population parameter is the mean):

  $\text{CI} = \bar{x} \pm z * \frac{s}{\sqrt{n}}$

  The Z score for 95% confidence is 1.96

- It applies to other parameters as well, we are looking at the mean just as an example...
- Let's do it on SPSS with a different dataset...
- Calculate the confidence interval for "age";
- What does this result tell us?

Confidence Intervals

- *Limits constructed such that for a certain percentage of samples (eg.: 95%) the true value of the population parameter will fall within these limits;*
- Confidence Interval (in this case the population parameter is the mean):

  $\text{CI} = \bar{x} \pm z * \frac{s}{\sqrt{n}}$

  The Z score for 95% confidence is 1.96
- It applies to other parameters as well, we are looking at the mean just as an example...
- Let's do it on SPSS with a different dataset...
- Calculate the confidence interval for "age";
- What does this result tell us?

Confidence Intervals

- *Limits constructed such that for a certain percentage of samples (eg.: 95%) the true value of the population parameter will fall within these limits;*

- Confidence Interval (in this case the population parameter is the mean):

  $\text{CI} = \bar{x} \pm z * \frac{s}{\sqrt{n}}$

  The Z score for 95% confidence is 1.96

- It applies to other parameters as well, we are looking at the mean just as an example...
- Let's do it on SPSS with a different dataset...
- Calculate the confidence interval for "age";
- What does this result tell us?

Confidence Intervals

- *Limits constructed such that for a certain percentage of samples (eg.: 95%) the true value of the population parameter will fall within these limits;*

- Confidence Interval (in this case the population parameter is the mean):

  $\text{CI} = \bar{x} \pm z * \frac{s}{\sqrt{n}}$

  The Z score for 95% confidence is 1.96

- It applies to other parameters as well, we are looking at the mean just as an example...
- Let's do it on SPSS with a different dataset...
- Calculate the confidence interval for "age";
- What does this result tell us?

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.

One Sample T-test

Compare the mean of a continuous variable to a specified constant value, 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);

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

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
  - $r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$
  - $-1 \leq r \leq 1$

Work in pairs

1. Let's go back to analysing the conflict dataset on SPSS;
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. Present your results to your classmates.

References

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