lab1-policy-analysis (2023)

December 7, 2023

1 Lab 1: Immigration policy chellenges | Justice and fairness Issues

As part of this work, you will conduct a descriptive and comparative analysis of the data, aimed at establishing relations between the assessments of people's opinions and socio-demographic and other characteristics. Imagine that you are entrusted with the task of a colleague from the policy planning department who would like to record a certain situation in relation to migration policy and work with citizens, and also separately consider the problem of perception of a sense of justice among residents of different countries. You need to work with *"historical"* data in order to identify some initial settings from different countries. At the same time, you are given an important condition: you are offered specific 2 countries for comparison for waves 1 and 9. The tasks for waves 1 and 9 are different, therefore, to obtain the maximum score, you must complete two blocks of analysis using two sources of sociological data. The list is presented in a separate Excel file.

The job organization is as follows:

- You need to work in RStudio and MS Word (LibreOffice Writer). In RStudio, you need to create the "Surname___Lab1" project, which is subsequently packed into a ZIP archive and attached to the answer as a Word document (Writer). The template for the answer is provided to you in a separate file.
- You will have two blocks of questions, each block has a maximum of 5 questions. Different questions are devoted to different aspects. In each answer it is necessary to give a comment (detailed answer) from 150 words to 450(650) words. A higher assignment score can be obtained by including tables and/or figures in addition to text. The quality of tables and images are also assessed, be careful!
- Each task is worth 0.5 to 2.5 points. A gradation scale for the quality of the response is provided from 0 to 1 with a step of 0.25. If something was slightly incorrectly described or attached as a table or figure, -0.25 points are taken off inside the task. The assessment of the assignment also includes the quality of writing code in the R language. If there are errors or missing steps, -0.25 points are deducted depending on the level of the problem. The final grade is the sum of accumulated points, rounded arithmetically.

1.1 Sources of literature:

1. [Task 1] Monogan IIId's book. Chapters 3 & 4 + 5.2
2. [Task 2] Monogan IIId's book. Chapter 3.
3. [Task 3] Monogan IIId's book. Chapter 5.1 + Handout 1+2
4. [Tasks 7-8] Monogan IIId's book. Chapters 5.3 + Handout 1+2

1.2 Part 1. Analysis of migration problems in countries.

Problem situation: There is a social problem related to a question how the social environment perceived by citizens and immigrants changes, how citizens are willing to accept them into their cultural, social and economic spaces, and which can be largely characterized by people's attitudes towards migrants. There is always a request for state-of-the-art study concerning such topics, with an accent on the social-demographics and other aspects, f.e., concerning attitudes towards migrants.

The purpose of the study: to describe the value and relational profile of the inhabitants of the countries proposed to you for analysis, from the standpoint of socio-demographic, migration characteristics and citizenship.

Technical details: Dataset label in Rstudio must be formed as `essw1`. You will need to create a single micro-database containing only two countries provided to you individually (see Excel file). You can use template Script in order to outline those actions. Those ones are to be included as part of the Script. Here is a template sctipt for that purpose:

[ ]:
```
## Importing data for Part 1: ESS Wave 1
library(haven)
ESS1e06_6 <- read_sav("ESS1e06_6.sav")
View(ESS1e06_6)

# Working with Cntry1 and Cntry2 and call the dataset as essw1
essw1 <- subset(ESS1e06_6, cntry == "##" | cntry == "##")
# Remove original dataset
rm(ESS1e06_6)
gc() # collect any garbage and put to trash can
```

1.2.1 Task 1 [2 points]:

Formal Task: *Please provide colleagues with a descriptive analysis of the sample structure in terms of socio-demographic characteristics. Compare specifications of those attributes between countries and in general. Describe similar and different trends.*

Task Description: Find socio-demographic variables in the questionnaire (either in PDF or in Excel - sheet ESSW1). Find the same variables (their labels) in the resulting (reduced to two countries) array `essw1`. Use known steps (functions) in order to study their structure of answers (scales). If the variables can be considered as *nominal or categorical*, create copies of the specified variables in the same dataset, which will be converted to the `factor` format [`factor()`] function. Please carefully do this handling for all variables. Try to find exact or similar variables in *Section F*. You are recommended to study Codebook for Wave 1 to get variable label in microdata & list of options. List of required variables to check:

- Gender of the respondent
- Age of the respondent
- Respondent's education
- Employment status
- Category of financial situation

- Place of residence of the respondent

Write code to build a set of descriptive statistics for each of the variables. In the document, write a comment describing the share distribution (*Percentage*) of *nominal/categorical variables* in general and by country in particular + *general tendencies (mean/sd/max/min/quartiles) for numeric/continuous* ones. In RStudio, when writing a function, you can build a crosstable of a series of frequencies in order to systematize all the necessary characteristics. You can also create one or more tables for the document and a series of plots, in which such a description will be presented. Your final target is to charcterize the selected questions in terms of general tendencies (and 2 countries combined) for these aspects. When performing cross-tables, please work only with table output. The corresponding plot called *matrix plot* is not informative. If you want not to see them, in function `crosstab(...)` you can include such an option `plot = FALSE` according to this template:

```
crosstab(dataset$variable_of_interest, dataset$grouping_variable, prop.c = TRUE, plot = FALSE)
```

Check-list:

1. All nominal variables are setup as `factor` value types. Numeric/continuous ones are left `as is`
2. You understand the idea of frequency tables and contingency tables. You provide evidence only in terms of shares in %.
3. You remember that for nominal and categorical variables you can plot barplots or pies. For nominal/continuous/7-10Likerts – histograms and boxplots.

!NB: when plotting boxplots in groups, as well as statistical testing per groups → *variable* = numeric/continuous one, and *category* is a nominal/categorical (f.e. *country*). Hence you see such a template in boxplot.

[ ]:
```
# Some useful template examples:
dataset$nominal_variable <- factor(dataset$original_variable, levels = c(val1,␣
↪val2, valk), labels = c(val1_lab, val2_lab, val3_lab))
# 2 plots in side-by-side
par(mfrow = c(1,2))
hist(dataset$var1, ...)
hist(dataset$var2, ...)
# When copied, press "tidy" button

# For boxplot in comparison with 2 countries [template]
boxplot(dataset$var3 ~ dataset$cntry, ...)

# Plotting area 2x2
par(mfrow = c(2,2))
hist(dataset$var1, ...)
hist(dataset$var2, ...)
hist(dataset$var3, ...)
hist(dataset$var4, ...)
# When copied, press "tidy" button
```

1.2.2 Task 2 [1.5 points]:

Formal Task: *Please formulate general and country-specific opinions and attitutes about their integration into country's environment you are provided to study.*

Task Description: Use `Part D` of the questionnaire. Find any of *6 varibles* in such a way that it should be **2** variables with 5-scales (*agree/disagree*) and **4** numeric/continuous/Likert-scale-based. When working with them please follow some of the recommendations below:

- For 5-scales variable please create a `factor` variable with option `ordered = TRUE` [`factor(dataset$var_factor, levels = ... , labels = ..., ordered = TRUE)`]. Two of 5-scaled variables are to be used further for Task 4. But you select at this stage.
- When dealing with numeric/continuous/Likert-scale-based please make sure you eliminate technical missings (refusals, don't know and so on). You can eliminate them by assigning such values to `NA`.

Write code to prepare those variables according to recommendations above. Then make a series of histograms for the 2 variables you are dealing with in terms of numeric/continuous/Likert-scale-based the following way:

- Prepare a 2x2 area with `par()` function to plot 4 histograms and/or boxplots at once. First row will consider 2 variables out of the whole sample (2 countries altogether, whole sample). Second row depicts the same variables, but subsetted by particular countries chosen for you.
- In order to plot country-specific values, please study central tendencies (sd, mean, median, min/max, IQ & IIIQ) of the 4 variables you have chosen, and provide interesting option. Again, variables for boxplot/hist must be only numeric/continuous, as they are to be used in Task 3 further.

In a document please insert the single boxplot and provide comments on the central tendencies for the variables you have chosen as well as comment on the boxplot in general and between countries in details.

Check-list:

1. You know how to identify 5-scaled 'agree/disagree' ones. You have found them in a questionnaire as well as in the dataset.
2. You know how to provide summaries for binary/nominal/catgorical variables and for numeric/continuous/7-10-Scaled. You know about *mean*, *median*, *standard deviation*, *quartiles* and *min/max*
3. You know that if you have missing codes, they have to be eliminated → converted to `NA`.
4. You know how to setup the 2x2 plotting area and include the plot sequently.
5. You know how to make nice-looking histograms and/or boxplots
6. You know how to plot a histogram/barplot/boxplot for specific country or other condition.

[ ]:
```
# Examples for Check-List point No. 3
dataset$var1_newlabels <- car::recode(dataset$var1, "some_value = NA") #␣
 ↪variant 1
dataset$var1_newlabels[is.na(dataset$var1)] <- NA                      #␣
 ↪variant 2
```

1.2.3 Task 3 [1 point]:

Formal Task: *Please provide functional check for normality criterias of the variables you have chosen. Describe the statistical tests results and choose further strategy between paramentic-based and nonparametric-based tests*

Task Description: Continue working with the variables you've chosen. Select those ones which are applicable for normality testing. Using 2 functions for normality testing (Anderson-Darling and Shapiro-Wilk), assess all of those ones. In the document you are to write about the variables you are going to assess using such tests (Remember that a statistical test deals with normality. Then think about proper sets of H_0 and H_1). Create a table with all the technical information to be provided by those tests. The layout can be the following:

- Column 1: variable/question names (or wording of the variables)
- Column 2: Anderson-Darling Test
- Column 3: Shapiro-Wilk test

!NB: Variables for boxplot/hist must be only numeric/continuous.

In a document write a comment about the results provided. For each of the variables being processed write a H_0 hypothesis and alternative H_1. Provide evidence about which of them are you going to use for parametric/nonparametric testing and why (how it can be proved)?

Check-list:

1. You know about the variable types applicable for the normality test
2. You can identify the H_0 and H_1 for such tests
3. You can relate the output with the interpretation and further steps

1.2.4 Task 4 [2.5 points]:

Formal Task: *Please provide functional analysis of differences in means/medians for numeric/continuous variables depending on agreement questions chosen, some of the social-demographic characteristics and country specifics. Give conclusions about the differences in those measures*

Task Description: Now you will need to create *such a copy of the dataset* labelled as `essw1_small` which contains only those variables you've chosen and prepared for the Lab. Find them in the end of the variable list using `names(essw1)`. Use template `essw1_small <- essw1[c('var1', 'var2', ...)]` in order to save them. Now work with `essw1_small`. You will need to make a series of parametric or non-parametric tests for *comparing means* or *comparing medians*. Please provide such a series using the following accents:

1. migration variable (option 1) & gender categories
2. migration variable (option 2) & country categories
3. migration variable (option 3) & employment status categories
4. migration variable (option 4) & place of residence of the respondent categories
5. respondent age & employment status categories

For points 2, 3 and 5 above provide a series of boxplots, which measure central tendencies across the variables and categories. Make a custom plotting area of `3x1` style. Provide meaningful titles, xlabs and ylabs, color scheme. In the Document please write a detailed comment about the results of the

compare means/medians test. Prepare a single table which include all the technical information about each pair of variables/categories for a test.

---

2 Part 2: Justice and fairness Issues

The purpose of the study: to describe the value and relational profile of the respondents in terms of *timing of life* and *political/income fairness*. .

Technical details: Dataset label in Rstudio must be formed as `essw9`. You will need to create a single micro-database containing only two countries provided to you individually (see Excel file). You can use template Script in order to outline those actions. Those ones are to be included as part of the Script. The procedure is the full copy of the case with `essw1`. You continue in the same script as for Part 1.

2.0.1 Task 5 [0.5 points]:

Formal Task: *Choose the variables that deal with fairness of income and timing of life. You will need to select 10 variables. Describe their wording, type and content of options.*

Task description: Work with Section D and G. From each section choose any 5 variables that you think of being most suitable to study according to the country you were provided. In the Text Document please write a comment about your choice and suggest possible explanation.

2.0.2 Task 6 [1.5 points]

Formal task: *Please provide colleagues with a descriptive analysis of the sample structure in terms of socio-demographic characteristics. Compare specifications of those attributes between countries and in general. Describe similar and different trends.*

Task Description: Find socio-demographic variables in the questionnaire (either in PDF or in Excel - sheet ESSW9). Find the same variables (their labels) in the resulting (reduced to two countries) array `essw9`. Use known steps (functions) in order to study their structure of answers (scales). If the variables can be considered as *nominal or categorical*, create copies of the specified variables in the same dataset, which will be converted to the `factor` format [`factor()`] function. Please carefully do this handling for all variables. List of required variables to check:

- Age of the respondent
- Number of years of education gained (F16)
- Employment status
- Influence in employment (F27 or F28)
- Number of contracted hours of work (F29)
- 5 variables from Section D
- 5 variables from Section G

!NB: Social-demographic variables are always mandatory in any sort of research, hence you must inlude them as mandatory alongside with target varitables from Secion D & G. If you experience problems with creation of correlation matrix with `cor()`, then do not setup `factor` types for nomminal/categorical variables. Then calculated matrix will not contain NA. That means you need just to select only variables for `essw9_small`. If still something goes wrong in this case,

try to manually grant R possibility to tread values are pure numeric (even if they were nominal categories), using this step after such one dealing with creeation of `essw9_small`:

```
essw9_small <- apply(essw9_small, 2, as.numeric)
```

Do not forget to check for missings for each of the variables (NA's or negative values.) Make them pure NA's then. Write code to build a set of descriptive statistics for each of the variables. In the document, write a comment describing the share distribution (*Percentage*) of *nominal variables* in general and by country in particular + *general tendencies (mean/sd/max/min/quartiles) for numeric/continuous* ones. In RStudio, when writing a function, you can build a crosstable of a series of frequencies in order to systematize all the necessary characteristics. You can also create one or more tables for the document and a series of plots, in which such a description will be presented. Your final target is to charcterize the selected questions in terms of general tendencies (and 2 countries combined) for these aspects.

Check-list:

1. All nominal variables are setup as `factor` value types. Numeric/continuous ones are left `as is`
2. You understand the idea of frequency tables and contingency tables. You provide evidence only in terms of shares in %.
3. You remember that for nominal and categorical variables you can plot barplots or pies. For nominal/continuous/7-10Likerts – histograms and boxplots.

!NB: Useful template code from Session 7 for corrplot with p-values:

```
wvs_fc <- cor(wvs_factor, use = "na.or.complete", method = "pearson")
wvs_fpv <- cor.mtest(wvs_factor, conf.level = 0.95)
corrplot(wvs_fc, p.mat = wvs_fpv$p, sig.level = 0.10,
         method = "number", col = terrain.colors(10))
```

2.0.3 Task 7 [1.5 points]

Formal task: *Please generate the correlation analysis (corrplot) for the variables selected for each of the country selected separately*

Technical description: You work with those variables from the previous task. You need to general a correlation matrix and a correlation plot using `corrplot(...)`. For that, generate a small part of the datased called `essw9_small`, the copying the proccess from Task 5. Work with this smaller dataset. Using `cor(...)` generate a correlation matrix. You can apply `as.matric(cor(...))`. Then load the package corrplot via `library(corrplot)`. Follow the recoomendations below. When generating a correlation matrix, please apply `use = ...` and `method = ...` options. You need to decide between Spearman and Pearson correlation. When you have generate the corrplot, copy-paste to the Document. You can also copy correlation matrix to the table. In the Document please provide a comment with relation strength analysis between the variables you have chosen. Relate with some theoretical thought about possible reasons for that relationship.

Check-list:

1. You know how to make matrix data type
2. You know how to make the correlation matrix as a matrix table

3. You know how to make smaller datasets via selection of particular variables out of the source one
4. You know how to visualize and interpret the correlation values according to p-values.

[ ]:
```
# install.packages("corrplot", dep = TRUE)
library(corrplot)
essw9_cor <- as.matrix(cor(essw9_small, use = "....", method = "...."))  # make
 ↪correlation matrix
essw9_pv  <- cor.mtest(essw9_cor, conf.level = 0.95) # make the p-values
corrplot(essw9_cor, p.mat = essw9_pv$p, sig.level = 0.1, method = 'pie', "....")
```

2.0.4 Task 8 [1 point]

Formal task: *Provide final reflections of the analysis you have provided (>100 Words). What are the main hightlights you can accent on out of the analysis?*

Technical description: You work with the Document. final reflections of the analysis you have provided. What was the most interesting in your output?