Political Analysis Using R

Descriptive Statistics

Hilary 2016

Authors: Roosmarijn de Geus, Niels Goet, and Julia de Romémont

In this lab we will focus on working with actual political data and answering descriptive research questions. The topic we will focus on in the lab sessions is women's representation in politics. We will use the dataset created by Arend Lijphart and used in his book *Patterns of Democracy*. The dataset contains information about women's representation in parliament and in the cabinet in 1990 and 2010 and covers 36 countries.

In this lab we will focus on descriptive statistics, which can reveal a lot of interesting information. Examples of descriptive research questions are what the average level of women's representation in Europe is in 1990, how it has changed between 1990 and 2010, which countries perform best or worst, and whether there is a lot of variation in women's representation across various countries. This lab provides you with the skills to answer such questions on a basic level.

1. Working with real data:

- First things first, if we work with an existing data-set we need to read this data into `R` by using the `read()` function, `R` can handle a wide variety of data-sources such as `STATA` or `SPSS` or `EXCEL` files, as well as files that are stored online
- We will use the Lijphart data-set which is a textfile where the values are separated by comma's, a `csv` file
- Open a new R-Script and use the code below to read in the Lijphart dataset. Note that we are assigning the data the name 'data', using the `<-` operator

```
data <- read.csv("http://andy.egge.rs/data/L.csv")
```

2. Exploring the dataset

Let's explore the dataset by using some of the commands you learned last week

```
head(data) #first six rows
tail(data) #last six rows
names(data) #variable names
str(data) #gives the structure of the dataset and variable types
data #shows the whole dataset
```

The dataset is saved as a data frame, and we can access variables in a variety of ways, for instance using the `object[row,column]` or the `$` notation. Some are more useful than others. Try out the options below:

```
data[1,1] #for the first row and the first column
data[1:5,] #the first five rows for all columns
data[,1:5] #the first five columns for all rows
data$country # $ notation to call variables from the dataset
data[1:4, c("country","hdi_2010")] #view multiple variables
```

Try this out for yourself and answer the following questions:

- What is the name of the 35th variable in the dataset?
- What is the name of the 8th country in the dataset?
- What is the voter turn-out for the first six countries?
- What is the unemployment rate for 1981-2009 and 1991-2009 for Denmark, Finland and France?
- How much has the foreign aid budget of Luxemburg increased between 1990 and 2005?

3. Research questions: descriptive statistics

Now that you have gotten a feel for the data, let's delve into some more interesting descriptive research questions! As said before, we are going to focus on women's representation in politics. In this lab, we'll focus on representation in parliament. The variables are `womens_parl_representation_1990` and `womens_parl_representation_2010`. First, let's make our lives easier and change the variable names:

```
names(data)[55] <- "women1990"
names(data)[56] <- "women2010"
```

3.1 Central tendency

A good first step is to look at various measures of central tendency such as the `mean` or the `median` level of women's representation for the countries in the dataset. Look these up for the year 1990 using the code provided below:

```
mean(data$women1990)
median(data$women1990)
summary(data$women1990)
```

- Which countries perform best and worst in 1990?
- How does this change in 2010?

We can dig a bit deeper and look at the quantiles, as well as tell `R` to give us the value of women's representation for the *lowest* 10% of the countries, or the *highest* 10%

```
quantile(data$women1990)
quantile(data$women1990, c(0.1, 1))
```

A different way of exploring the level of women's representation is to order the countries based on their scores. Try it out:

```
data$country[order(data$women1990)]
data$country[order(data$women1990, decreasing=TRUE)]
```

- Which country has the highest percentage of women in parliament? And which country the lowest?

3.2 Variability and distribution of data

Two final measures we can look at are the `variance` and the `standard deviation`. Both are measures of variability, they provide information about the distribution of your data. Specifically, they inform you to what extent your data is clustered around the `mean`, and whether your data has a very wide distribution:

- The `variance` is the sum of squared differences of the observations from the mean
- The `standard deviation` is simply the square root of the variance

Both formulas are provided below, as well as their R codes. Try them both out in R to check that you get the same answer.

Variance formula

$$variance = s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2 \tag{1}$$

```
# Variance by hand from formula in R:
sum((data$women1990 - mean(data$women1990))^2)/ (length(data$women1990)-1)

# R Function:
var(data$women1990, na.rm = TRUE) # variance
```

Standard Deviation Formula

$$s.d. = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_i (x_i - \bar{x})^2} \tag{2}$$

```
# Standard deviation by hand from formula in R
sqrt(sum((data$women1990 - mean(data$women1990))^2)/ (length(data$women1990)-1))

# R Function
sd(data$women1990, na.rm = TRUE) # standard deviation
```

Now have a closer look at these measures and answer the following questions:

- What is standard deviation for women's representation in 1990? How about 2010?
- How do they differ? What does this tell you about women's parliamentary representation across the 36 countries?
- Do you find the standard deviation an informative measure in answering potential research questions? Why (not)?

4. Missing data

You might wonder what the `na.rm = TRUE` in the `R` functions displayed above mean... This relates to missing data - as a default `R` assumes there is no missing data, however in many datasets there will be (quite a few!) observations missing. The addition `na.rm = TRUE` tells `R` to ignore these when computing the statistics that you ask it to. It is easy to check whether your variable of interest has missing data:

```r
is.na(data$women1990) # a logical vector listing missing data
```

- Now try calculating the mean for the consumer price index 1981-2009 (`cpi_1981_ 2009`) using the `mean()` function. What happened?
- Check if the variable has missing data with the `is.na()` function.
- Now try telling `R` that your variable has missing data and to ignore it using the `na.rm = TRUE` option:

```r
mean(data$cpi_1981_2009, na.rm = TRUE)
```

5 Plots

Now that we have calculated descriptive statistics, let's use `R` to create some visuals:

```r
boxplot(data$women1990) # creates a boxplot of your data
boxplot(data$women1990, data$women2010) # adds 2 boxplots
boxplot(data$women1990, main = "My Title", xlab = "myxlabel", ylab = "myylabel")
# adds some nice labels and titles
```

Answer the following questions about the boxplot:

- What does the thick black line indicate?
- What do the whiskers represent?
- Which information is contained in the box?
- What do the dots mean?
- Based on these boxplots, what can you say about women's representation in parliament?

Another helpful way to visualise your data is a histogram:

```r
hist(data$women1990) # creates a histogram of your data

#Now as you can see, R has decided to provide you with a title
#and x-axis and y-axis labels. You can change these using
#the title = "", xlab = "", and ylab = "" commands shown above.
```

6. Correlation

Now that we have explored descriptive statistics, we can delve a bit deeper. As a final step we will look at associations and correlations between two variables. We might think for instance that there is an association between women's representation in parliament and other variables in the dataset.

Have a look at the variables in the dataset and write down potential options:

- _____
- _____

- ---

We might think for instance that countries that have higher levels of development as measured with the human development index (`hdi_2010`) also have a higher number of women in parliament. We can visualise this association as follows (note that we use the representation 2010 variable (`women2010`)!):

```
plot(data$hdi_2010, data$women2010) # creates a scatterplot
text(data$hdi_2010, data$women2010, labels = data$country, cex = 0.7)
# the command on the second line adds the country labels
# the "cex = 0.7" option specifies the font size of the country labels
```

Excellent! Your first scatterplot! Now try and interpret the plot:

- Based on the plot, how would you describe the association between human development score and women's representation in parliament?

Luckily, we can ask `R` to calculate a statistic of the strength of the association between two variables: the correlation coefficient. The `cor.test()` command tests for correlation between two variables:

```
cor.test(data$hdi_2010, data$women2010)
```

- What is the correlation coefficient? What does this tell you?
- Is the correlation statistically significant? How can you tell?
- Try flipping the two variables around and re-run the correlation test. Do you find any difference? Why (not)?

We can visualise the correlation with a regression line (more on regression in lab 3!):

```
plot(data$hdi_2010, data$women2010) #note: here, it's first the x and then the y variable!
abline(lm(data$women2010~data$hdi_2010)) #note: here, it's first the y and then the x variable!
```

7. Homework

Deadline: Friday of 3rd week, noon.
Send your answers to the following two questions (including `R` code used) to your lab instructor:

- Create a scatterplot that shows human development index (`hdi_2010`) on the x-axis and representation of women in parliament in 2010 (`women2010`) on the y-axis.
- What is the correlation between these two variables? Is the correlation statistically significant?

Main commands used in this lab session

```
# data[row,column]
# data[1:4, c("var1", "var2")]
# mean(); median(); summary()
# percentile()
# data$var1[order(data$var2)]
# var() and sd()
# is.na()
# boxplot(); hist(); plot()
# cor.test()
```