Introduction to Statistics
lab 1

Johan A. Elkink
jos.elkink@ucd.ie

7 September 2015

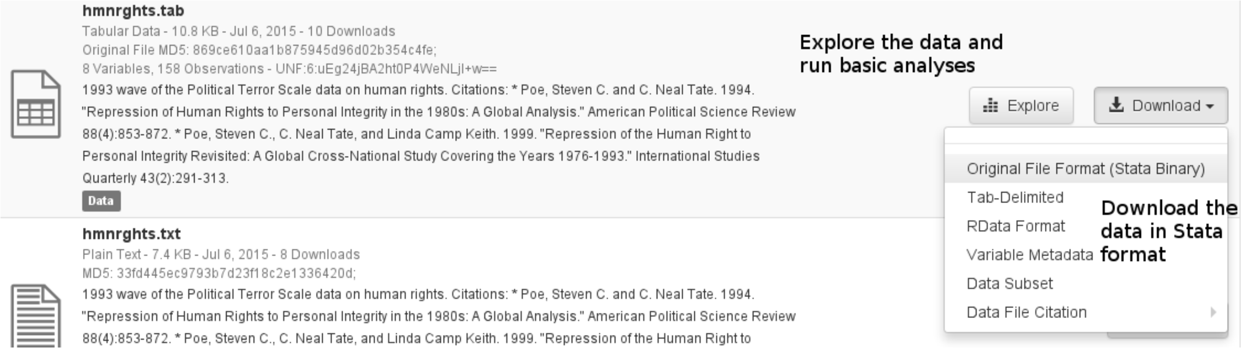
The main purpose of today's class is to get a feel for how to open, access and view data, and to get some familiarity with the user interface of the software.

If you are proceeding fast, you might want to try doing the same exercise in several software packages, to get an idea of how they operate.

Downloading the data

The first task is to download the data we will use for this lab from the web. We will make use of data from Poe, Tate and Keith (1999), which has been made available for class exercises by Monogan (2015). This data is made available through the Harvard Dataverse, a data repository that contains a large number of social science data sets, including all replication data from a number of prominent journals in political science.

Go to `https://dataverse.harvard.edu/` and search for "political analysis in r", then go to the data repository of Jamie Monogan. Download the data file `hmnrights.tab` in Stata format (Figure 1).



hmnrghts.tab
Tabular Data - 10.8 KB - Jul 6, 2015 - 10 Downloads
Original File MD5: 869ce610aa1b875945d96d02b354c4fe;
8 Variables, 158 Observations - UNF:6:uEg24jBA2ht0P4WeNLjl+w==
1993 wave of the Political Terror Scale data on human rights. Citations: * Poe, Steven C. and C. Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis." American Political Science Review 88(4):853-872. * Poe, Steven C., C. Neal Tate, and Linda Camp Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A Global Cross-National Study Covering the Years 1976-1993." International Studies Quarterly 43(2):291-313.

hmnrghts.txt
Plain Text - 7.4 KB - Jul 6, 2015 - 8 Downloads
MD5: 33fd445ec9793b7d23f18c2e1336420d;
1993 wave of the Political Terror Scale data on human rights. Citations: * Poe, Steven C. and C. Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s: A Global Analysis." American Political Science Review 88(4):853-872. * Poe, Steven C., C. Neal Tate, and Linda Camp Keith. 1999. "Repression of the Human Right to

Explore the data and run basic analyses

Download the data in Stata format

Figure 1: How to download data from DataVerse

Note that on the same page you can also click on "explore" to give you a basic description of the data. On the left-hand side you will find a list of all variables and by hovering your mouse over the variable name, you will find a series of descriptive statistics.

Opening the data

Open the data you downloaded in your preferred statistics package.

SPSS: Open SPSS (PASW). Go to File, Open, Data and find the file you downloaded. Make sure you select "Stata" under file types, otherwise you will not see the file. Click on the filename. Make sure to use "Paste" instead of "Ok" when you are ready to open the file.[1] This will generate the code to open the data but not actually take any action just yet. Go to the Syntax screen of SPSS, select the lines of code that were created, and click on the green play button to execute the command. This should open your data. Have a look at the Data and Variable screens. Add a line of comments in the syntax file explaining what data set you're opening. Save the syntax file for your convenience. For the remainder of the lab exercise, include everything in the syntax file and keep saving it as you go along.

R: Open RStudio. In R, it is generally good practice to always write the code yourself – there is no "Paste" button as in the graphical user interface of SPSS. Go to File, New, R Script and a new blank R script editor window will open. R has a large number of libraries to extend its functionality – to open Stata files is one of those additional functionalities. So you need to start with opening the library using:

```
library(foreign)
```

which is the library to open foreign data sets. You can then open the data file using:

```
hr <- read.dta(file.choose())
```

The `file.choose()` command gives you a dialog screen to find the file and returns the name and full path of the file.[2] The `read.dta()` command opens Stata files, and the assignment operator `<-` is used to assign the file to a name in R. All objects in R need to have a name. We arbitrarily used `hr` in this case. Select the lines of code and press "Run". The variable `hr` should show up in the list of variables – double click on it to see the data. Add a line of comments in the R file explaining what data set you're opening. Save the R file for your convenience. For the remainder of the lab exercise, include everything in the R file and keep saving it as you go along.

Stata: Open Stata. Go to File, Open, Data and find the Stata data file that you downloaded. Open the data. On the left you will see a history of command run and a list of variables. Open a new do-file editor using File, New, Do. Click on the command in the history file and copy it to the do-file editor. This command should be something along the lines of:

```
use "C: ...  hmnrights.dta"
```

which is Stata command language for opening files. Click on Data, Data Editor to see the data.

[1] This is *always* true when "Paste" is available as an option.

[2] Try just running the `file.choose()` command without `read.dta`.

Add a line of comments in the do file explaining what data set you're opening. Save the do file for your convenience. For the remainder of the lab exercise, include everything in the do file and keep saving it as you go along.

Inspecting the data

Tables for one variable

You can produce a table of a single variable with the following code, for example for the **democ** variable:

SPSS: `FREQUENCIES democ.`

R: `table(hr$democ)`

Stata: `tab democ`

Try this with a number of different variables.

Obtaining the mean

We will discuss the mean in the next class (basically, the sum of a variable divided by the number of cases), but here is how you can obtain this for a variable:

SPSS: `DESCRIPTIVES democ.`

R: `mean(hr$democ, na.rm = TRUE)` – note that the `na.rm` parameter is required to avoid an error message that the variable contains missing data.

Stata: `su democ` – where `su` is an abbreviation of summarize.

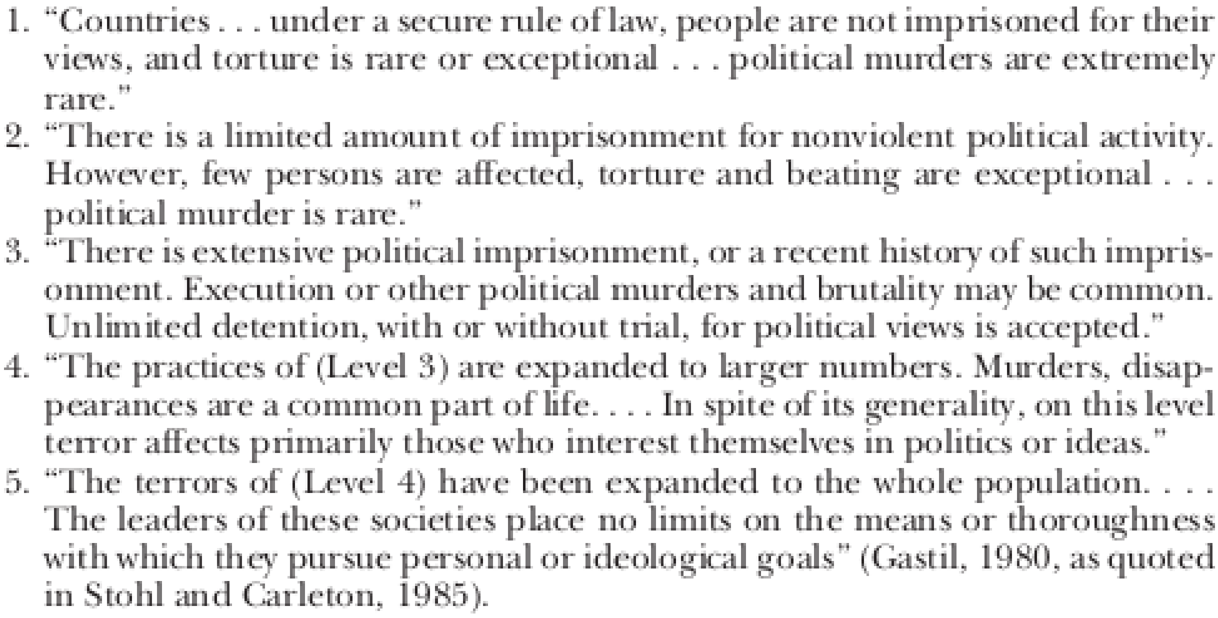
Try this with a number of different variables.

Pie charts

You are probably familiar with pie-charts. Here is how you create a pie chart of the **gnpcats** variable, which is a categorized version of the Gross National Product measure for each country:

SPSS: `FREQUENCIES gnpcats /PIECHART.` – note that here pie charts are an optional parameter to frequency table.[3]

[3] An alternative is `GRAPH /PIE = gnpcats.` .



1. "Countries . . . under a secure rule of law, people are not imprisoned for their views, and torture is rare or exceptional . . . political murders are extremely rare."
2. "There is a limited amount of imprisonment for nonviolent political activity. However, few persons are affected, torture and beating are exceptional . . . political murder is rare."
3. "There is extensive political imprisonment, or a recent history of such imprisonment. Execution or other political murders and brutality may be common. Unlimited detention, with or without trial, for political views is accepted."
4. "The practices of (Level 3) are expanded to larger numbers. Murders, disappearances are a common part of life. . . . In spite of its generality, on this level terror affects primarily those who interest themselves in politics or ideas."
5. "The terrors of (Level 4) have been expanded to the whole population. . . . The leaders of these societies place no limits on the means or thoroughness with which they pursue personal or ideological goals" (Gastil, 1980, as quoted in Stohl and Carleton, 1985).

Figure 2: Categorization of the measure of violations of personal integrity. Source: Poe, Tate and Keith (1999: 297)

R: `pie(table(hr$gnpcats))` – note that here the pie chart function takes as parameter a frequency table. Note also the use of the dollar symbol to access variables (here `gnpcats`) within data sets (here `hr`).

Stata: `tab gnpcats, generate(f)`, followed by `graph f1 f2, pie` – note that Stata pie charts make use of the output from the tabulate command (here saved as new variables starting with "f").

In all cases, pie charts and tables are closely connected. How do you think frequency tables and pie charts are related?

Cross-tables

The main variable of interest in Poe and Tate (1994); Poe, Tate and Keith (1999) is the measure of violations of personal integrity in different countries. Figure 2 gives an overview of the classification, which in the data is named as **sdnew**. Lets investigate whether countries under military control (**military** equals 1) are more prone to violate personal rights, using basic tabulation.

SPSS: `CROSSTABS /TABLES = sdnew BY military.`

R: `table(hr$sdnew, hr$military)`

Stata: `tab sdnew military`

Generally it will be helpful to have percentages here instead of just raw counts, since there are far fewer countries under military rule than there are not under military rule.

SPSS: `CROSSTABS /TABLES = sdnew BY military /CELLS = COLUMN.` – alternatively you can use "ROW" to get row percentages.

R: `prop.table(table(hr$sdnew, hr$military), 1)` – alternatively you can use 2 to get column percentages.

Stata: `tab sdnew military, row` – alternatively you can use "column" to get column percentages.

Some additional tools

SPSS: `CODEBOOK.`

R: `str(hr)`, `summary(hr)`, `head(hr)`.

Stata: `list`

Questions

Based on trying different variations of the above commands, try to answer the following questions:

1. What are the dimensions of the data set: how many cases are there? How many variables?
2. What is the unit of analysis in this data set?
3. What is the level of measurement for each of the variables?
4. Do you see any missing data? Has this affected any of your tables or figures?
5. What is the average (mean) population size? *(trick question!)*
6. What is the proportion of countries that are in civil war?
7. How long is the scale for the democracy measure? Name a highly democratic and a highly undemocratic country according to this data set.
8. Are military regimes more prone to violations of personal integrity according to this measure?

References

Monogan, Jamie. 2015. "Political Analysis using R: Example code and data, plus data for practice problems.".
http://dx.doi.org/10.7910/DVN/ARKOTI

Poe, Steven C. and C. Neal Tate. 1994. "Repression of Human Rights to Personal Integrity in the 1980s: A global analysis." *American Political Science Review* 88(4):853–872.

Poe, Steven C., C. Neal Tate and Linda C. Keith. 1999. "Repression of the Human Right to Personal Integrity Revisited: A global cross-national study covering the years 1976–1993." *International Studies Quarterly* 43(2):291–313.