SPSS syntax for descriptive statistics

Johan A. Elkink

February 28, 2013

1 Opening data

First, you need to open the data file. This might be an SPSS file with all the data properly defined, or it can be a file in another format that requires a bit more work. Most example data in this course is saved in Stata's `.dta` format, which is useful because it can be easily opened in a variety of statistics packages, but it does require a bit more manipulation to be fully usable.

To open a Stata file in SPSS, you can use:

```
GET STATA FILE = 'asiabaro.dta'.
```

Often you will need to specify the location of the file, for example:

```
GET STATA FILE = 'c:\Users\11020101\Downloads\asiabaro.dta'.
```

Or you can use the menus by using "File", then "Open", then select `.dta` files, then browse to the file you need, and then click "Paste".

After that, you select this command, and run it using the green arrow button. This should open the data file.

1.1 Selecting cases

For the example below on democratic and development, we want to select only one year from the sample, e.g. 1990.[1] You can use the following code here:

```
GET STATA FILE = 'demdev.dta'.
SELECT IF (year = 1990).
```

2 Univariate graphical descriptives

Graphs in SPSS can be generated in two different ways: using what is called the "legacy dialogues", which is the menus that were common in older versions of SPSS, or the "Graph

[1] In the example we look at the variable `laggdppc`. What this measures is the Gross Domestic Product per Capita, lagged by one year, i.e. the observations in the data set for 1990 are really the GDPpc levels of 1989. This handout ignores that detail.

Builder". The latter is more flexible, but the syntax code it generates is much more complicated. In this handout, I will use the old-fashioned code, but you are free to use either. In the Graph Builder, it is primarily a matter of dragging plot types and variables to the center screen and then click "Paste" to generate the syntax code.

For a pie chart, you can use:

```
GRAPH /PIE = religion.
```

This will create a pie chart of the `religion` variable in the `asiabaro.dta` data set, which represents religious denomination, a nominal variable.

If you get problems because the variable is not properly defined as nominal, you can use:

```
VARIABLE LEVEL religion (NOMINAL).
```

For the bar plot, you can use:

```
GRAPH /BAR = religion.
```

For scale level variables (interval or ratio measurement level), the histogram is more useful, for example using GDP per capita in the `demdev.dta` data set:

```
GRAPH /HISTOGRAM = laggdppc.
```

Or you can use a box plot:

```
EXAMINE laggdppc /PLOT boxplot.
```

Computing logarithms

For "money variables", such as GDP per capita, it is often useful to look at the logarithmic transformation. This transformation "stretches" the range of lower values and "squeezes" the range of higher values, such that the distribution of a variable that is very skewed with a long tail on the higher values, will approximate more closely a normal distribution or bell curve. You can calculate such variable and reproduce the plots with:

```
COMPUTE logGDPpc = ln(laggdppc).
GRAPH /HISTOGRAM = logGDPpc.
EXAMINE logGDPpc /PLOT boxplot.
```

3 Univariate numerical descriptives

The frequency table, which just lists the number of occurrances of a particular category for a categorical variable, can be generated with:

```
FREQUENCES VARIABLES = polity2.
```

This is always a good starting point to look at your data, but for scale variables, the table will be unwieldy.

To calculate the mode, median, and mean, you can use the following code:

```
FREQUENCIES VARIABLES = polity2
  /FORMAT = NOTABLE
  /STATISTICS = MEAN MEDIAN MODE.
```

Here, the NOTABLE option will suppress the actual frequency table, and the listed statistics will then be printed. Note that the EXAMINE command used to produce the box plot, also included these statistics in a table before the plot. Similarly, you can use this command to get statistics on variation, including the range, variance, and standard deviation:

```
FREQUENCIES VARIABLES = polity2
  /FORMAT = NOTABLE
  /STATISTICS = RANGE VARIANCE STDDEV.
```

4 Multivariate descriptives

Two scale variables

Graphically, the relation between two scale variables can be shown using a scatter plot, for example with:

```
GRAPH /SCATTERPLOT = laggdppc WITH polity2.
```

Note that the first variable (here laggdppc) will be in the x-axis and the second variable on the y-axis – typically, we put the dependent variable on the y-axis.

Numerically, the parallel to the univariate variance is the covariance, which can be calculated with the same command as the correlation. To get both, the following command works:

```
CORRELATIONS /VARIABLES = polity2 laggdppc democracy
  /STATISTICS XPROD
  /MISSING = PAIRWISE.
```

Related to correlation, we can also run a simple (only one independent variable) regression, with:

```
REGRESSIONS
  /DEPENDENT polity2
  /METHOD = ENTER laggdppc.
```

A scale and a categorical variable

Graphically, one approach to study this would be separate boxplots for the dependent variable, for each category on the independent variable. For example, so compare the distribution of democracy scores between countries with civil wars and those without, you could use:

```
EXAMINE VARIABLES = polity2 BY cwar
  /PLOT = BOXPLOT
  /NOTOTAL.
```

Two categorical variables

Here a cross-table is the most common visualisation of the relationship, whereby a key decision is the correct calculation of the percentages. They should be calculated over the categories for the independent variable, so that you can compare across the categories of the dependent variable. Example code would be:

```
CROSSTABS /TABLES = cwar BY democracy
  /CELLS = COUNT COLUMN.
```

or

```
CROSSTABS /TABLES = cwar BY democracy
  /CELLS = COUNT ROW.
```

Recoding a variable

Often, you might want to recode a variable into fewer or different categories. Here is the example as used in the exercise in the slides:

```
RECODE polity2
  (MISSING=SYSMIS)
  (Lowest thru -7=1)
  (-6 thru 6=2)
  (7 thru Highest=3)
INTO regime.
```

Note that you should always also add proper labels for both the variable and the respective values, and set the right level of measurement:

```
VARIABLE LABELS regime "Political regime classification".
VALUE LABELS regime
  1 "Autocracy"
  2 "Anocracy"
  3 "Democracy".
VARIABLE LEVEL regime (ORDINAL).
```

And finally, check whether it worked the way you expect:

```
CROSSTABS /TABLES = polity2 BY regime
  /CELLS = COUNT.
```

Saving a standardized variable

z-scores of a variable can be saved using:

```
DESCRIPTIVES VARIABLES = polity2 laggdppc /SAVE.
```

This will create two new variables, `Zpolity2` and `Zlaggdppc`.