Introduction to Statistics
lab 2

Johan A. Elkink
jos.elkink@ucd.ie

17 September 2018

Downloading the data

For this second lab we will use the replication data from Tajima (2013), which can be accessed on the course teaching data page[1] and is listed as "Indonesian census".[2] See the instructions for the previous class on how to open this file, which is also in Stata format. Remember to always save the file first, then open the statistical package of your choice, then open the file—double-clicking the file name does not work. Make sure you also open the description file, so that you have a list of variables in the data set, with brief descriptions.

For R users, I will assume you opened the data using "indo" as the data set name, i.e. `indo <- read.dta(...)`.

Note that while checking the correct syntax online in preparation for the lab sheets, I end up making a lot of use of the resources at `http://www.ats.ucla.edu/stat/`, which has excellent guidelines on coding in R, Stata, SPSS, and SAS.

Recoding

It is often necessary or helpful to recode variables. Sometimes we want to reduce a categorical variable into fewer categories (e.g. reducing all smaller categories to one "other" category) or reduce an interval or scale variable to a categorical one. Here we will see an example of the latter.

Recode the **population** variable into a small (less than 1200), middle (1200 to 3300) and large (greater than 3300) category. Note that we make sure to create a new variable (**popcat**) to avoid overwriting the original data.

SPSS:
```
RECODE population (lo THRU 1200=1) (1201 THRU 3300=2) (3301 THRU hi=3)
INTO popcat.
```

[1] `http://www.joselkink.net/data.php`
[2] Taken from the PODES DataVerse, `http://hdl.handle.net/1902.1/19477`.

```
VALUE LABELS popcat 1 'low' 2 'middle' 3 'high'.
```
—Note how we label the different categories to make the use of this new variable easier to understand.

R:
```
indo$popcat <- cut(indo$population, breaks = c(-Inf, 1200, 3300, Inf))
```

Stata:
```
gen popcat = population
recode popcat (min/1200=1) (1201/3300=2) (3301/max=3)
```

In all packages, procude a cross-table of the original variable **population** by the new categorical variable **popcat** to see if the transformation worked. You will also see why a scale variable like **population** is not well suited for tables!

Frequencies

Using code discussed in Lab 1, produce a frequency table and a pie chart for the new variable. What do you conclude about the distribution of this new variable?

Cross-tables

Using code discussed in Lab 1, produce a cross-table of **popcat** by **violence**, to see whether urban regions have lower or higher levels of intergroup or interethnic violence. What would be the correct percentages to include, row or column? Including the right percentages, what do you conclude?

Distributions

For continuous variables, we have to look at the data in a different way. Lets have a closer look at the original (before the transformation) **population** variable. First, calculate the mean, median, variance, and standard deviation of the **population** variable.

Numerical descriptives

SPSS:
```
DESCRIPTIVES population /STATISTICS = MEAN STDDEV VARIANCE.
FREQUENCIES /VARIABLES = population /FORMAT NOTABLE /STATISTICS ALL.
```

R:
```
mean(indo$population, na.rm = TRUE)
median(indo$population, na.rm = TRUE)
var(indo$population, na.rm = TRUE)
sd(indo$population, na.rm = TRUE)
```

Stata: `su population` —you might also want to try out `inspect population` and `codebook population`.[3]

Graphical distribution

Produce histograms and box plots for the **population** variable:

SPSS: `GRAPH HISTOGRAM = population.`
`EXAMINE population /PLOT BOXPLOT.` —Note that you could do both in one command, as in: `EXAMINE population /PLOT BOXPLOT HISTOGRAM.`

R: `hist(indo$population)`
`boxplot(indo$population)` —If you have plenty spare time, try `?hist` and `?boxplot` to find more options and parameters to tweak the presentation of the plot (title, color, labels, etc.).

Stata: `histogram population`
`graph box population`

Based on the graphical and numerical description of the variable, how would you describe the distribution of the **population** variable?

Why do you think the mean and median of the variable are so different?

Transforming the variable

In some cases where the distribution of the variable is highly skewed, it can be worthwhile to transform the variable. For example, when it comes to population or income variables, you will often have far more cases with low values than with very high values, and a logarithmic transformation can be helpful. Calculate the log transform of the **population** variable and repeat all of the above in the "Distributions" section for this new variable.

SPSS: `COMPUTE logpop = LN(population).`

R: `indo$logpop <- log(indo$population)`

Stata: `gen logpop = ln(population)`

How does the distribution change?

Extra

Only if you have plenty of time left over:

[3] See `http://homepages.rpi.edu/~simonk/pdf/UsefulStataCommands.pdf`

- Investigate, using the same tools, the distribution of the **inequality** variable.
- Look at the relationship between population size and inequality. Without transforming the variable, this would be:
  SPSS: `GRAPH /SCATTERPLOT = inequality WITH population.`
  R: `plot(inequality ~ population, data = indo)`
  Stata: `twoway (scatter inequality population)`[4]

References

Tajima, Yuhki. 2013. "The institutional basis of intercommunal order: Evidence from Indonesia's democratic transition." *American Journal of Political Science* 57(1):104–119.

[4] Perhaps also try `twoway (scatter inequality population), scheme(economist)`.