# **BQA WEEK 1**

## **EXERCISE 2:**

## **CROSS-TABULATIONS & SCATTERPLOTS IN STATA**

**+++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++**

DESCRIPTION: This exercise is split into two parts. First, you will draw on the skills covered in Exercise 1 to produce cross-tabulations, that include row or column percentages and chi-square values. Second, we cover how to produce single and two-way scatterplots to visually assess the relationship between continuous data.

FILES: **BQA\_Lab\_v2.dta**

### STEPS:

1. Open a new do.file titled ‘Exercise 2’ so you can save and replicate your work
2. Add explanatory comments as you go along
3. Tabulate discrete data and calculate chi-square values
4. Produce scatterplots
5. Add confidence intervals to the data
6. Save the do-file regularly

### NEW COMMANDS:

* Inspecting the data: **bysort**
* Analysing data: **tab, row, col, nofreq, chi, scatter, twoway, lfitci, qfitci**

### 2.1 Cross-tabulations (contingency tables)

The data set contains several variables which have been derived from other variables in the dataset. These appear at the bottom of your variable list, with names in capital letters (the original ESS variables all have lower-case names). These new variables have been “tidied up” in terms of having coded missing values set to system missing, and in many cases some of the categories have been combined, in order to make the crosstabs smaller.

The syntax for a cross-tabulation in Stata is:

**tabulate [1st var] [2nd var] [IF statements if required] ,[options]**

*Note: the square brackets are only there to separate out the arguments – you never type them in!*

Try tabulating MARSTAT (marital status) against LONELY, without any IF statements or options. What happens when you change the order of the two variables?

|  |
| --- |
|  |

If you don’t specify any options, you’ll only get frequencies. As we discussed in the lecture, these are not all that useful – percentages are much more useful. Suppose you decide to put MARSTAT as the first argument, and LONELY as the second. If we are interested in how the incidence of loneliness varies by relationship status, should we ask for row or column percentages?

|  |
| --- |
|  |

CHECK your answer to this question before moving on!

Experiment with the options **,row** and **,nofreq** and fill in the table below

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **No option** | **,row** | **,nofreq** | **,row nofreq** |
| **What do you get as output?** |  |  |  |  |

### 2.2 Chi-square statistic

You can get a chi-squared statistic for this cross-tabulation by including the option **,chi**

Include this option and fill in the table below, using an IF statement to examine the two subsamples of those who are 70+ and under 22. Consult the do-file for the appropriate syntax if you get stuck.

|  |  |  |
| --- | --- | --- |
|  | **p-value for chi-squared statistic** | **Degrees of freedom** |
| **Whole sample** |  |  |
| **Those aged  70 or over** |  |  |
| **Those aged  under 22** |  |  |

Why do you think you observe different results for the different subsamples?

|  |
| --- |
|  |

### 2.3 Row or columns percentages?

Let’s now look at the relationship between left-right political ideation (LRSCALE2) and attitudes to same-sex relationships (FREEHMS).

A couple of questions to ensure that you understand how to interpret row and column percentages. Use the full sample, specifying row or column percentages as appropriate.

Please take some time (after the class if necessary) to make sure you understand these answers.

|  |  |
| --- | --- |
| **What percentage of people who define themselves as “neither leftwing nor rightwing” agree strongly that gays and lesbians should be free to live their lives as they wish?** |  |
| **What percentage of people who disagree strongly with this proposition define themselves as very right-wing?** |  |
| **What percentage of the whole sample define themselves as very right-wing?** |  |
| **Given that an individual defines him- or herself as leftwing (but not very leftwing) what is the probability that he or she will agree, or agree strongly, with the proposition in question?** |  |

In the previous example, when we looked at marital status and loneliness, we found that the effects varied substantially between subgroups. Here, we’ll look at variations between countries. We’ll write our commands in the form **tabulate FREEHMS LRSCALE2,** in order to get some practice in interpreting column percentages.

### 2.4 Sorting tabulations by country

We could look at individual countries one at a time, by using IF statements. But here’s a handy little command (strictly, it’s a prefix) which will allow you to perform this function for all the countries together (you’ll need to click on “more” several times to get all the output, then you can scroll up and down between the countries).

**bysort COUNTRY: tab FREEHMS LRSCALE2,** **col nof chi**

You’ll see that the p-values for the chi-squared statistic indicate that the relationship is statistically significant in every single country. However, the nature of the relationship differs from country to country. Fill in the following table:

|  |  |  |  |
| --- | --- | --- | --- |
| **Countries in which there is NO significant relationship (p>0.05), or where it’s hard to discern the direction of the relationship** | **Countries in which (very) left-wing people are more favourably disposed towards same-sex relationships than (very) right-wing people** | **Countries in which (very) right-wing people are more favourably disposed towards same-sex relationships than (very) left-wing people** | **Countries in which both very left-wing and very right-wing people tend to have very definite views on same-sex relationships** |
|  |  |  |  |

Is there a pattern here? What might it indicate?

|  |
| --- |
|  |

### 2.5 Scatterplots

For an illustration of the “beehive” problem, try typing:

**scatter GovSat AGE**

(which plots the satisfaction with government scale against individuals’ ages). I think you’ll agree it’s impossible to deduce anything at all from this scatterplot! There’s simply too much information.

This problem will apply to pretty much anything we try to plot using the whole data set. However, the data set contains some derived variables which measure various factors at country level. There is only one observation per country (try using the **summarize** command to see this).

|  |  |
| --- | --- |
|  | **Derived from individual-level variable:** |
| **IMPRICH** | imprich (the importance of being rich) |
| **IMPCREATE** | Ipcrtv (the importance of being creative) |
| **LS** | LifeSat (life satisfaction scale) |
| **GETBY** | hincfel (feelings about getting by on household income) |

The syntax for straightforward scatterplots is pretty easy:

**scatter [var1] [var2]**

Have some fun trying it out on the variables above, and fill in the table below with + or – signs, indicating the direction of the relationship (some of the cells are greyed out so you don’t plot a variable against itself, and so you don’t do the same thing twice).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **IMPRICH** | **IMPCREATE** | **LS** | **GETBY** |
| **IMPRICH** |  |  |  |  |
| **IMPCREATE** |  |  |  |  |
| **LS** |  |  |  |  |
| **GETBY** |  |  |  |  |

Labelling the points is also pretty easy – it takes the form of an option, in which you specify the variable that you would like to use to as a marker for the names. So, for example, to indicate that we want to label the points using the country variable, type in:

**scatter IMPRICH IMPCREATE, mlabel(COUNTRY)**

Describe the pattern that you see emerging.

|  |
| --- |
|  |

### 2.6 Two-way scatterplots

Finally, we’d like to add a line of best fit to the graph. In Stata this is a little tricky, but well within our capabilities. You have to plot TWO graphs, one on top of the other, one with the scatterplot and the other with the line. There are rather a lot of parentheses, but you should be able to work out what they all mean.

**twoway (****scatter IMPRICH IMPCREATE, mlabel(country)) (lfit IMPRICH IMPCREATE)**

You may like to substitute qfit for lfit, which allows for a little bit of curvature in the relationship.

**twoway lfit IMPRICH IMPCREATE** (and **twoway qfit IMPRICH IMPCREATE**) will draw the line (and quadratic line) of best fit, without the scatterplot.

This gives us a potential solution to the “beehive” problem.

Try **twoway** **lfit** and **twoway qfit** for the relationship between satisfaction with government and age (**GovSat** and **AGE**). What do the results tell you?

|  |
| --- |
|  |

Finally, you may like to try **lfitci** and **qfitci**, which also provide confidence intervals.