POLS 7012

Ryan Bakker

UNIVARIATE STATISTICS AND GRAPHS

Topic: Graphical presentation of data. Univariate statistics, including mean, mode, median, standard deviations, percentiles *etc*. Tabulating and summarizing data.

STATA commands and features: `tabulate`, `summarize`, `menu-driven graphs`

Data set: wvs.dta, taken from the World Values Survey 1991.

More information: http://www.worldvaluessurvey.org/services/index.html

Readings: Alan Agresti and Barbara Finlay. 1997. *Statistical Methods for the Social Sciences, 3rd ed.* Upper Saddle River, NJ: Prentice Hall. [CHAPTER 3].

T.H. Wonnacott and R.J Wonnacott. 1990. *Introductory Statistics*, New York: Wiley. [CHAPTER 2].

1. GRAPHICAL PRESENTATION OF DATA

One of *STATA*'s strengths is that it is a relatively powerful graphics program. Graphs can be created through both syntax and menu-driven commands. Although throughout this course we focus on syntax commands, it is often easier to draw graphs using the menu-driven commands. This is because the **graph** syntax command has many options (listed in the *STATA* online help facility). There are enough options, in fact, that it is virtually impossible to graph from the command line without the online help description. Given this, we will use the menu-driven option for specifying graphs. The menu itself generates *STATA* commands, which you can then modify if you like (*i.e.*, although we run the command through the menu, the syntax appears in the command window). We will demonstrate some of *STATA*'s graphing capabilities using the religiosity variable (*religious*) in the WVS.

Perhaps the simplest exploratory graph is the histogram, which by default expects continuous variables; so we can look at age by:

Graphics
Histogram
Variable
*Choose the variable to graph (age)
OK

Note that if you click 'Submit' (rather than 'OK') the tab-sheet remains active, so you can amend and resubmit without waiting for it to reappear.

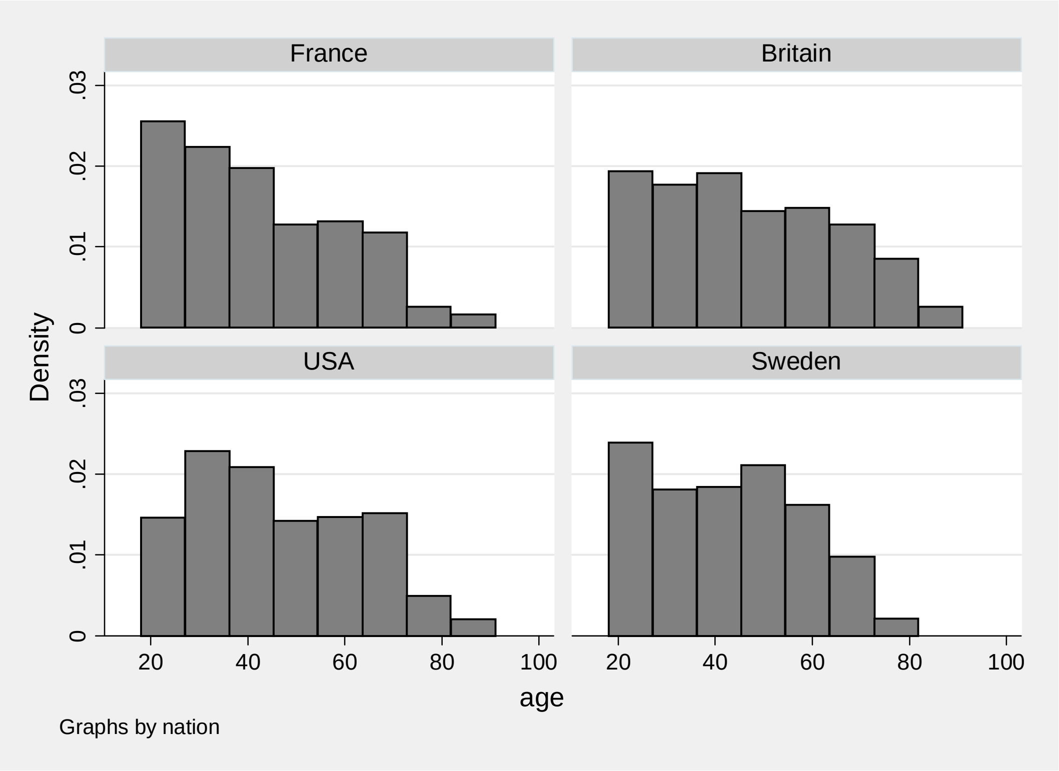
We can alter our graph by using the options within the tabs at the top of the **histogram** menu. For example, we can change the level of detail provided by the histogram by shifting the number of 'bins':

Main tab
Number of Bins
* Choose number of bins (8)

We can also look at histograms across some other variable using the **by** tab:

By tab
By controls
Variables
* Choose 'by' variable (nation)

You should get a graph that shows the varying age distributions across countries:



You can use a histogram to look at categorical variables too (such as *religious*), but as the default is for continuous variables, you should warn *STATA* with the **discrete** modifier:

Main tab
Variable
* Choose religious
* Mark 'Discrete data'
OK

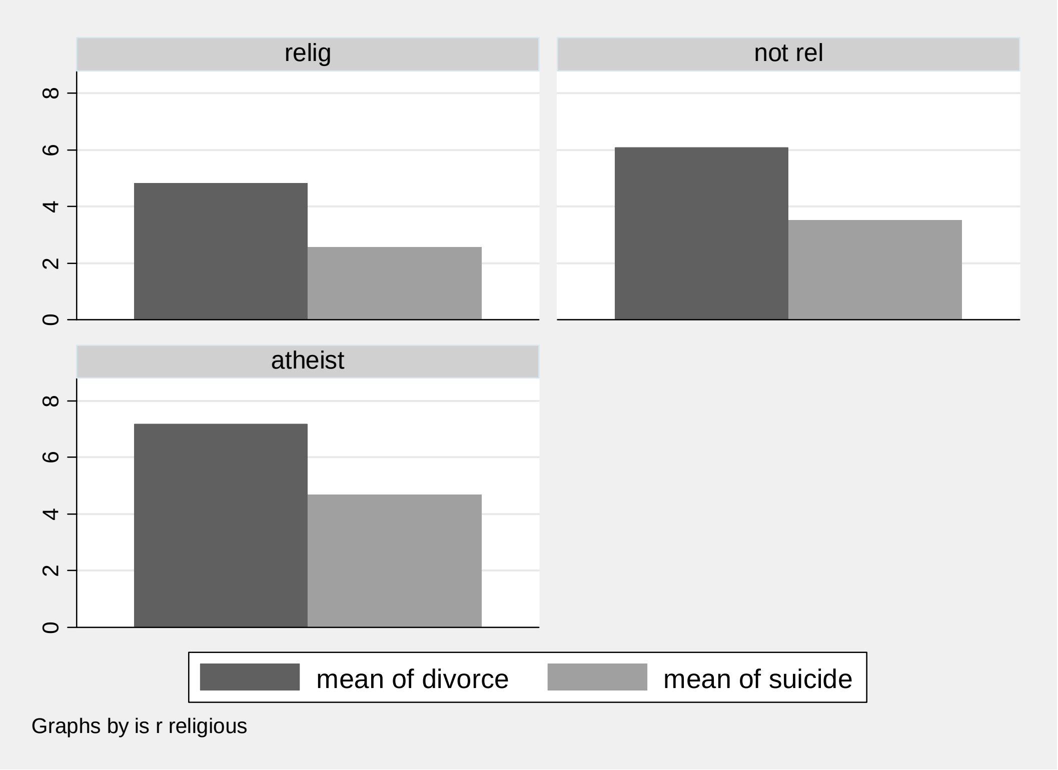
The scale of histograms is often immaterial; but most familiar for categorical variables would be scaling by proportions, which the default scaling achieves.

It is also easy to add titles through the menu tabs: **`y-axis`**, **`x-axis`** and **`titles`**.

Whilst the tidied-up histograms are more elegant than those produced by the minimalist command, notice that the analytic information has not increased. When exploring and trying to make sense of data, the short forms of the graphing command are often adequate. However, if your graphs are to be seen by others (*e.g.* prepared for publication, or indeed the class assignments), *you must make sure that they are presentable*. Good presentation aids understanding: it is not an optional extra.

The last histograms we drew for discrete variables were effectively operating as bar charts. *STATA* also has an explicit bar chart option, which reports by default the mean values of the variables listed. If we were interested in differences in attitudes to divorce (*divorce*) and suicide (*suicide*) by religiosity we might do:

Graphics
Bar charts (Summary statistics)
Statistic ‘Mean’, Variables
*Choose the variables to graph (divorce suicide)
By tab
By controls
* Choose religious
OK



Graphs by is r religious

Once again, you can tidy up the titles if you wish.

We can use the **over groups** tab instead of the **by** tab to produce a different layout of the same information:

- Graphics
  - Bar charts (Summary statistics)
    - Statistic 'Mean', Variables
      - *Choose the variables to graph (divorce suicide)
        - Over Groups tab
          - Over 1
            - * Choose religious
              - OK

For our present purposes, the **by** and **over groups** tabs affect layout, not substance. To discover means by variable1 and variable2 we simply add another variable name in the relevant box.

Following the logic of the command, if we wish to use the **bar charts** menu to give us the proportions in each category of 'religiosity' that we need to make dummy variables corresponding to each category and display their mean values. So first we have to make the dummy variables. We can do this either step by step, as:

```
. recode religious (3 = 1 atheist) (nonm = 0 other), into(relig3)
. recode religious (2 = 1 nonreligious) (nonm = 0 other), into(relig2)
. recode religious (1 = 1 religious) (nonm = 0 other), into(relig1)
```

or more simply:

```
. tab religious, gen(relig)
```

We then deploy the resulting three variables in the graph command, thus:

- Graphics
  - Bar charts (Summary statistics)
    - Statistic 'Mean', Variables
      - *Choose the variables to graph (relig1 relig2 relig3)
        - Over Groups tab
          - Over 1
            - * Choose nation
              - OK

As we said at the beginning of this section, *STATA* has extensive options to modify the layout of its analysis graphs. These can be best explored through playing around with the menu commands, as well as the **help** menu, and we shall meet some of them in later sessions.

EXERCISE 1

Following on from last week’s exercises, create some graphs to illustrate the relationship between age and religiosity (is age or our created variable age-group most suitable here?). Use graphical methods to show whether the relationship varies by country.

Experiment with the options available in the menus to add titles, change colours etc.

2. SUMMARIZING CATEGORICAL DATA

Univariate statistics come in two forms that most people look at: measures of central tendency and measures of dispersion. The simplest type of univariate statistic is a measure of central tendency, this indicates how the typical or average person behaves or what the typical or average value of a variable is. We also care about how much a variable varies. If it did not vary there would be nothing to explain. So we calculate a measure of variation that tells us how dispersed cases are from our measures of central tendency.

Remember which measure of central tendency we use will depend on whether we have what is called a categorical or discrete variable as opposed to an interval level variable. Many of the variables in the WVS are categorical, which is true of many public opinion surveys; *e.g.*, yes, no, maybe. The numbers attached to each of these responses do not have any real meaning, which usually means that we cannot use the mean or median (which are mathematical calculations) to summarize these variables. Instead we can look at frequencies, the mode, or sometimes we might be able to create a 0-1 dummy variable and then use the mean as a summary. Generally speaking, most of the WVS variables are best summarized with a simple table,

. tabulate *variable*

using the **tabulate** command that we used last week. *peoptrust* is a good example: “Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?” [(1) Most people can be trusted, (2) Can't be too careful].

. tabulate peoptrust

```
     people |
    trusted |      Freq.     Percent        Cum.
------------+-----------------------------------
    trusted |      1,637       47.70       47.70
    careful |      1,795       52.30      100.00
------------+-----------------------------------
      Total |      3,432      100.00
```

If we wanted to use the mean as a summary, we could do the following:

Recode a new variable to preserve *peoptrust*

. recode peoptrust (1=1 "Trusting") (2=0 "Not Trusting"), gen(trust)

```
. summarize trust

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       trust |      3432    .4769814    .4995426          0          1
```

We know that about 48% of all respondents are trusting. The standard deviation is not particularly helpful; since all values are either 0 or 1, the standard deviation is a straightforward function of the sample proportion. The frequency table also gives you the mode, which is the category that occurs most frequently. If you want to do this by country the command is

```
. tabulate nation, summarize(trust)
```

EXERCISE 2

Now we will look at national pride. Create a new binary (0-1 or dummy) variable, pride, by recoding *proud* into a new variable where 1 = proud and 0 = not proud. What would a suitable division of people into two categories here be? Remember to label the new variable.

Then use the summarize command to give you an idea of how much national pride respondents have. Use the tabulate command with the summarize option to compare means in both trust and pride across countries.

3. SUMMARIZING INTERVAL DATA

For interval-level variables, there are additional univariate statistics that might be more useful, such as the mean, median and standard deviation. One example is *lifesat*: “All things considered, how satisfied are you with your life as a whole these days?”. Respondents give an answer from one to ten, where 1 is ‘dissatisfied’, 10 is ‘satisfied’ and 99 is ‘don’t know’. (In truth, this variable is not interval-level data. But many social scientists would use it as such, so we will follow suit). There are basically no interval level variables in this dataset, aside from age.

```
. tabulate lifesat

       life |
satisfactio |
          n |      Freq.     Percent        Cum.
------------+-----------------------------------
   dissatis |         23        0.67        0.67
          2 |         20        0.58        1.25
          3 |         63        1.84        3.09
          4 |         81        2.36        5.45
          5 |        294        8.57       14.02
          6 |        296        8.62       22.64
          7 |        574       16.72       39.36
          8 |        963       28.06       67.42
          9 |        591       17.22       84.64
     satisf |        525       15.30       99.94
         dk |          2        0.06      100.00
------------+-----------------------------------
      Total |      3,432      100.00
```

The tabulate command allows you identify the modal and median categories relatively easily. If we want additional univariate statistics (of both central tendency and dispersion), the easiest thing to do is use the summarize command:

```
. summarize [varlist] [if exp] [ , [detail]
```

With no options specified, the command gives you the mean, standard deviation, minimum and maximum. If you add the detail option, you will also get percentiles, variance, skewness, and kurtosis.

For *lifesat*, we would type:

```
. summarize lifesat if lifesat < 11, detail
```

Note that we have included the **`if lifesat < 11`** to exclude the two people that said "don't know" (or dk). Otherwise, the value 99 is included in all the univariate statistics.

```
                     life satisfaction
-------------------------------------------------------------
      Percentiles      Smallest
 1%            2              1
 5%            4              1
10%            5              1       Obs                3430
25%            7              1       Sum of Wgt.        3430

50%            8                      Mean            7.61312
                        Largest       Std. Dev.      1.815298
75%            9             10
90%           10             10       Variance       3.295308
95%           10             10       Skewness      -.8797813
99%           10             10       Kurtosis       3.793747
```

Percentiles give a reasonable idea of the distribution of the variable. We might also use a histogram:

Graphics
Histogram
Variable
*Choose the variable to graph (lifesat)
If/in tab
If: (expression)
* Type **`lifesat < 11`**
Density plots tab
Tick 'Add normal density plot'
OK

Notice the new option we have used with the **Histogram** menu: '**Add normal density plot**' adds the normal curve to compare the variables distribution to that of the normal distribution.

EXERCISE 3

The literature on social capital is closely allied with a body of literature dealing with ‘well-being.’ Life satisfaction is a central variable in this body of literature, as is the variable *decision*:

“Some people feel they have completely free choice and control over their lives, while other people feel that what they do has no real effect on what happens to them. Please use this scale where 1 means ‘none at all’ and 10 means ‘a great deal’ to indicate how much freedom of choice and control you feel you have over the way your life turns out”.

a) Find the modal and median categories as well as the mean. Create a similar graph to the one above. How does it compare to life satisfaction?
b) Using a categorical measure of age, examine whether positioning on the ‘decision making’ scale differs across different age groups. Create a graph that summarizes the relationship between age and scale position across the whole dataset.
c) Do we find the same relationship described in (b) in all countries in the dataset?