Practicum 1: Data description and graphics

Harvard Catalyst Certificate in Applied Biostatistics

Goal: This practicum will go over how to calculate summary statistics and how to make simple graphics in STATA.

Worked example

**Summary statistics**

In lecture, we investigated how to calculate summary statistics using the *summarize* command and how to produce basic graphics using the *histogram, graph box* and *twoway*. In this practicum, you will use the same commands to investigate the dataset.

Please load the dataset into STATA by clicking the dataset link on the webpage. STATA should open automatically and the dataset should be loaded.

To calculate the summary statistics for a continuous variable, the STATA command *summarize* can be used. Let’s investigate age as we did in lecture.

Type the command: *summarize age*

The following output is produced:

. summarize age

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

age | 252 45.70635 10.50105 22 73

In this initial output, STATA provides the number of observations in the dataset and the mean, standard deviation, minimum and maximum age.

Type the command: *summarize age, detail*

The following output is produced:

. summarize age, detail

age

-------------------------------------------------------------

Percentiles Smallest

1% 25 22

5% 28 24

10% 31 25 Obs 252

25% 38 25 Sum of Wgt. 252

50% 45.5 Mean 45.70635

Largest Std. Dev. 10.50105

75% 53 69

90% 60 70 Variance 110.272

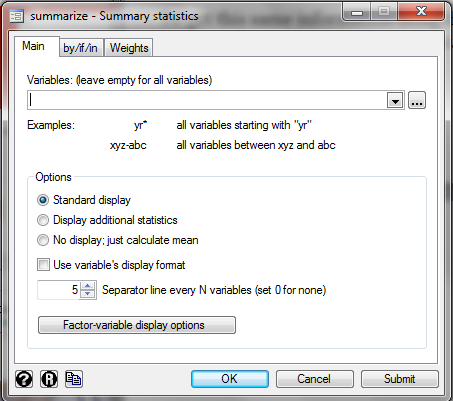
95% 63 71 Skewness .0715598

99% 70 73 Kurtosis 2.524414

In this output, STATA provides the additional statistics related to age.

To get this same information using the drop down menus, you use the following menus: *Statistics/Summaries, tables and tests/Summary and descriptive statistics/Summary statistics*

The following menu will open:



In this menu, you type “age” in the variables window. If you click OK, you will get the first output listed above. If you click “Display additional statistics” and then click OK, you will get the second output listed above.

**Graphics**

STATA has the ability to create many graphics. To see the entire list, please look at the options under the *Graphics* tab. The three graphs focused on in lecture were the histogram, box plot and scatterplot.

For the histogram of age, we use the following command:

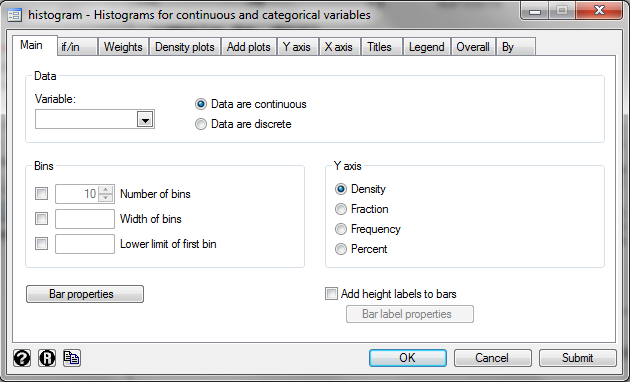
*histogram age, frequency xtitle(Age) title(Histogram of age)*



For this command, “age” is the variable name for the variable of interest. The frequency option tells STATA to have the y-axis be the number of people in each bin, which is my preference (the default is different). The other two options tell STATA the title of the x-axis and the title of the graph.

To create this graph using the drop down menus, you use the following menus: *Graphics/Histogram*

The following menu will open:

**

In this menu, you type “age” in the variable window. To get the frequency on the y-axis, you need to click the Frequency box above. To add labels to the X-axis and to add a title, you must go to the appropriate tabs in this menu.

For the box plot of age, we use the following command:

*graph box age, ytitle(Age) title(Box plot of age)*



Alternatively, you can use the following menu: *Graphics/Box plot*

For a scatterplot of age compared to mental quality of life score (mcs),

*twoway (scatter mcs age), ytitle(MCS) xtitle(age)*

**

Alternatively, you can use the following menus: *Graphics/Twoway graph (scatter, line, etc)*. Then, click the *Create* box to create a graph. In the resulting window, you place “mcs” in the y-axis box and “age” in the x-axis box. You click Accept and then click OK to create the scatterplot. Note that many alternative graphs can be created using this menu so please try on your own if you are interested.

**Proportion**

To calculate the proportion of males in the dataset, many STATA commands can be used, and we will cover some of these in future practicums. In the dataset, male=0 for females and male=1 for males. One command that provides a lot of valuable information is tabulate:

*tabulate male*

. tabulate male

male | Freq. Percent Cum.

------------+-----------------------------------

0 | 194 76.98 76.98

1 | 58 23.02 100.00

------------+-----------------------------------

Total | 252 100.00

This output provides the number of females (194) and males (58), the percent of females (77%) and males (23%), and the cumulative percentage. The cumulative percentage is the total percentage of all categories above.

Try on your own

Using the commands listed above, calculate the mean, median, maximum and minimum of the variable “pcs”, which a measure of Physical Quality of Life.

Create a histogram for “pcs” and label the axes.