**Statistical Thinking**

**Problem Set 1**

Due Sunday 24th at midnight – please email finished problem sets to Vianney Gomezgil Yaspik at [vg349@cam.ac.uk](mailto:vg349@cam.ac.uk).

**TYPES OF VARIABLES**

The table below lists five variables which are found in the data set. Inspect the variables using the **tabulate** command, and write in the table which type of variable each represents.

|  |  |
| --- | --- |
| **Description and name of variable** | **What type of variable is it?** |
| 1-10 Favourability Scale towards Angela Merkel (*like\_angela\_merkel*) |  |
| Capital City (*capital*) |  |
| Country (*country*) |  |
| Percentage who believe that “Global warming is a hoax” (*Glob\_conspiracy\_\_5*) |  |
| Region (*region*) |  |

**MORE ON THE TABULATE COMMAND**

Use the tabulate command to answer the following questions:

|  |  |
| --- | --- |
| What percentage of regions in this dataset are capital city areas? (variable capital) |  |
| How many regions in this dataset are located in Denmark? (variable country) |  |

**SUMMARY STATISTICS, USING THE SUMMARIZE COMMAND**

For the three continuous variables in the table below, find the statistics and fill in the table.

You will find the mean, median and SD using the summarize command with the “detail” option

(eg **summarize sm\_use\_facebook, d**)

The **mean** command gives the mean, standard error of the mean, and the 95% confidence interval for the mean (eg **mean sm\_use\_facebook**).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable** | **Median** | **Mean** | **Standard deviation** | **SE of Mean** | **95% CI** |
| Users of Twitter (sm\_use\_twitter) |  |  |  |  |  |
| Belief that “Regardless of who is officially in charge of governments and other organisations, there is a single group of people who secretly control events and rule the world together” (*Glob\_conspiracy\_\_4*) |  |  |  |  |  |
| Percentage who like Indian food (like\_indian\_food) |  |  |  |  |  |

If you really enjoy this sort of thing, you might like to check

1. that the 95% confidence interval really is equal to the mean plus or minus 1.96 standard errors.
2. That the standard error of the mean is equal to the sample standard deviation divided by the square root of the sample size.

**SORTING AND RANKING**

You can use gsort to sort the dataset before tabulating, in order to get a ranked list by that variable.

1. Which region has the highest favourability towards Aung San Suu Kyi (*fav\_assk*)?

2. In which *countries* are people most likely to believe that “humans have made contact with aliens and this has been kept secret” (Glob\_conspiracy\_\_6)?

**HISTOGRAMS FOR CONTINUOUS VARIABLES**

Histograms are created using the **histogram** command in Stata, specifying the name of the variable you want to use.

**hist earthsun** will produce a very smooth looking histogram, with a (fairly) steady distribution. Tabulating the **earthsun** variable shows that there are 69 values in the distribution. We could give each one its own bin, and use 69 bins:

**hist earthsun, bin(69)**

Adding **normal** as an option compares the distribution of the variable to the normal distribution.

**hist earthsun, bin(69) normal**

You don’t have to use 69 bins: 15 bins would produce a smoother distribution, since the categories are “doubling up”.

Draw histograms for the following variables, and fill in the table below:

|  |  |  |
| --- | --- | --- |
| **Description and name of variable** | **How well does it fit the normal distribution?** | **Positive or negative skewness?** |
| Favourability towards Barack Obama (fav\_obama) |  |  |
| Liking (1-10) rating for Vladimir Putin (like\_putin) |  |  |
| % who like sushi (like\_sushi) |  |  |

**BOX PLOTS**

Box plots are pretty easy in Stata. The syntax is (for example) **graph box fav\_pope**

You can do box plots of several variables together, should you wish: **graph box fav\_pope fav\_obama fav\_trump**

Stata box plots show the median; the 25th and 75th centiles; and the end of the whiskers extend the 25th and 75th centiles by 1.5 x the interquartile range (or, by the most extreme value within that range).

Make a box plot for the **fav\_gates** variable (worldwide favourability towards Bill Gates). Looking at the outlying values, would you say they indicate (a) slightly atypical regions; (b) hugely atypical regions; or (c) errors in data collection? Why?

|  |
| --- |
|  |