**Statistical Thinking**

**Problem Set 1**

**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*) | **Continuous.** Note this is not categorical or ordinal, because even though the underlying question is on a 1-10 scale, the variable in the regionally aggregated dataset is an average. |
| Capital City (*capital*) | **Dichotomous/dummy/binary.** This is a variable coded as 0 or 1. It is not the name of the capital city, so it is not categorical. |
| Country (*country*) | **Categorical.** |
| Percentage who believe that “Global warming is a hoax” (*Glob\_conspiracy\_\_5*) | **Continuous.** |
| Region (*region*) | **Categorical.** (As some people observed, there’s no region variable in this version of the dataset though there is a MATCHID variable). |

**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) | **16.5%** |
| How many regions in this dataset are located in Denmark? (variable country) | **1** |

**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) | **30.6** | **30.5** | **13.2** | **1.20** | **28.1 – 32.85** |
| 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*) | **25.3** | **26** | **10.3** | **.952** | **24.13 – 27.90** |
| Percentage who like Indian food (like\_indian\_food) | **21.1** | **28.3** | **21.8** | **1.99** | **24.39 – 32.25** |

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.

**Some people actually did this and showed their calculation (if you were one of those, kudos) – and pointed out that due to rounding, they do not 100% match (but are close)**

1. 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*)?

**Southeast Anatolia (TR7) – shorthand name, “Kurdistan” (Turkish)**

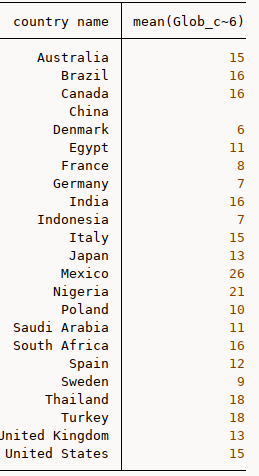
2. In which *countries* are people most likely to believe that “humans have made contact with aliens and this has been kept secret” (Glob\_conspiracy\_\_6)?

**The answer to this question is** **Mexico. A lot of people got the correct answer, but by a bad method – looking at regional averages, instead of country averages. This is unreliable, as just because a country has some of the highest regions on a particular item, this does not mean the average for the country *as a whole* is necessarily the highest.**

**To get the country mean average for all regions in that country, use the table command:**

**encode country, gen(nation)**

**table nation, stat(mean 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) | **Fairly well.** | **Positive** |
| Liking (1-10) rating for Vladimir Putin (like\_putin) | **Quite poorly.** | **Very weakly positive.** |
| % who like sushi (like\_sushi) | **Not too bad.** | **Positive** |

**(Well done to those who pasted in the histograms as well as pointing out the method of assessing skewness and fit to the normal distribution.)**

**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?

|  |
| --- |
| **Box Plot:**    **There is one outlier, Lagos. It is hugely atypical, with a value almost 4 standard deviations from the mean.**  **An argument could be made that this is due to errors in data collection. We know that for this item of the survey, samples were much lower due to randomisation of the appearance of names on the list, and that Lagos is a challenging environment in which to conduct public opinion research. A priori, a value of 83% seems implausibly high.**  **However as many have observed, error alone is unlikely to be the whole story, with Bill Gates having recently visited Nigeria, made it a priority country for the Gates Foundation, and paid off a portion of Nigeria’s national debt. Bill Gates is likely to appeal to different constituencies of public opinion in Lagos: to some for his business success (and partnership with Aliko Dangote), to some for his charitable endeavours in the country, to some for his criticism of the Nigerian government for not doing more to invest in public health services.** |