**Statistical Thinking**

**Problem Set 2**

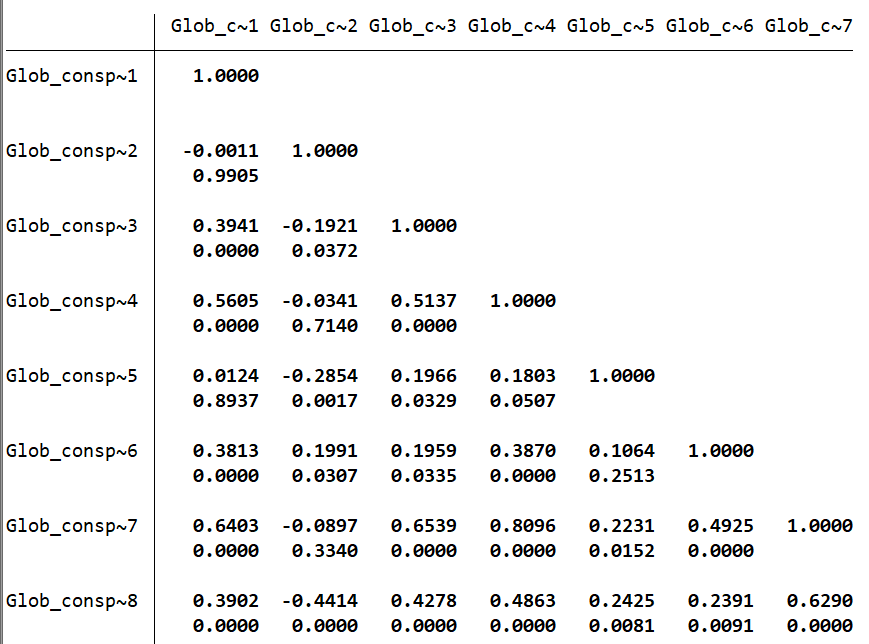
**CORRELATION COEFFICIENTS**

Your task is to create a correlation matrix of all the variables in the data set relating to conspiracy theories. These are the variables beginning **Glob\_conspiracy**

Now, you are ready to produce your correlation matrix

**pwcorr Glob\_c\***  will produce correlation matrix of variables starting “Glob\_c”

**pwcorr Glob\_c\*, sig** will also give p-values for every cell



|  |  |
| --- | --- |
| **One variable in the list appears to be less strongly correlated with the others. Which is it?** | **The variable with the lowest average absolute (i.e. positive or negative) correlation with the others is Glob\_conspiracy\_\_2 - which is also the only variable that is, on average, negatively correlated with the other items.**  **However Glob\_conspiracy\_\_5 was also an acceptable answer to this question, on the grounds that its average absolute (i.e. positive or negative) correlation with other items was almost (though not quite) as low as is the case with Glob\_conspiracy\_\_2.** |
| **Pick one correlation you find interesting, and use a scatterplot (command: scatter) with name labels (ml(short)) to discuss the cases you think are responsible for this association.** | **There was no right or wrong answer to this question as long as it was well discussed. Most people chose to scatterplot the items for belief that the world is run by a secret group, and belief that AIDS was spread by a secret group, on the grounds that both represent at least one common underlying conspiratorial worldview.** |

**ONE SAMPLE T-TESTS**

The syntax for conducting a one-sample t-test in Stata is

**ttest [variable] = [the value you are testing]**

For example,

**ttest earthsun = 75**

provides a test under which the null hypothesis is that the mean of earthsun is equal to 75 (i.e.: the regional average for the percentage of respondents who believe that “the earth goes around the sun” is actually 75% of the population, rather than the recorded mean of our sample, which is just 49%).

The top of the output gives summary statistics; below that, the null hypothesis is set out, and below that, three alternative hypotheses: that the actual mean is greater than, equal to, or less than 75.

Below these three hypotheses are the weight of the distribution in question.

Now test the following hypotheses, and briefly report your findings:

1. That across world regions, on average, a fifth (20%) of the population believe that “the truth about vaccines is being deliberately concealed” (**Glob\_conspiracy\_\_3**)

2. Across world regions, the average number of social networking sites that people make use of is exactly 4 (**n\_social\_media**)

**Note that for these questions, there was no need to report the alternative hypothesis results for a difference greater or less than zero – the questions only asks us to test the hypothesis that Glob\_conspiracy\_\_3 is 20 and n\_social\_media is 4 (for which the alternative is diff!=0).**

**(Unrelatedly, also note that when pasting Stata content in to Word, if you change the font to courier new it will align nicely on the page).**

**1. Stata output:**

**One-sample t test**

**------------------------------------------------------------------------------**

**Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]**

**---------+--------------------------------------------------------------------**

**Glob\_c~3 | 118 18.6671 .6758857 7.341998 17.32854 20.00566**

**------------------------------------------------------------------------------**

**mean = mean(Glob\_conspiracy\_\_3) t = -1.9721**

**Ho: mean = 20 degrees of freedom = 117**

**Ha: mean < 20 Ha: mean != 20 Ha: mean > 20**

**Pr(T < t) = 0.0255 Pr(|T| > |t|) = 0.0510 Pr(T > t) = 0.9745**

**There were many different ways of phrasing this answer, depending on whether respondents adopted the (rather stilted) language of frequentist statistics or a more down to earth explanation instead.**

**In more down to earth language -- we are being asked the probability that in the world as a whole, the average regional proportion who believe that “the truth about vaccines is being deliberately concealed” might really be 20% (rather than the 18.7% average of our sample).**

**The p value of 0.051 says that in a world in which the true value really *were* 20%, then we’d draw a sample like ours on about 1 in 20 (1 / 0.051) occasions.**

**That makes it fairly unlikely that we’d have gotten this data if the true value were in the region of 20%. Possible, yes - but fairly unlikely.**

**Now the formal answer…**

**The conventional hypothesis testing framework would have us fail to reject the null hypothesis at the critical threshold (alpha) of 0.05 (p < 0.05) based on a two-tailed test. The value of 20 lies within the 95% confidence interval. We thus fail to reject the hypothesis that the world regional average who believe that the “truth about vaccines is being deliberately concealed” is 20%.**

**Which is better? For policymakers, frankly I’d suggest the former is more informative. It could be in the region of 20%, but it probably isn’t. Who cares if the p value is 0.049 or 0.051? Either way, the chance of getting these results is about 1 in 20, assuming the true value were in fact 20%. The second formulation, while technically “correct”, actually obscures this fact, in a way that conveys a misleading degree of openness to the hypothesis that the true mean is 20%.**

**A final note:**

**Finally, bear in mind that strictly speaking what the results do not tell us is that “the chance of the true population mean being 20 is 5.1%”. Why not? Well, think about it – for any continuous variable the probability of the true mean being *exactly* 20.000000 has got to be vanishingly small – we cannot assign probabilities to exact point estimates. For statements about the probability of what the true population mean might be, we can only refer to ranges of values (e.g. “above 20%” or “below 20%”). And for that kind of statement, we have to use the one-tail test results. For example, the Ha mean > 20 result of 0.9745 tells us that, in a situation in which the true population mean was 20 or lower, we would get a sample like the one we happen to have, on 97.45% of occasions. Stated more intuitively and less formally, it means that there is an overwhelming probability that the true mean is less than 20%.**

**2. Stata output:**

**One-sample t test**

**------------------------------------------------------------------------------**

**Variable | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]**

**---------+--------------------------------------------------------------------**

**n\_soci~a | 121 3.859121 .093702 1.030722 3.673597 4.044644**

**------------------------------------------------------------------------------**

**mean = mean(n\_social\_media) t = -1.5035**

**Ho: mean = 4 degrees of freedom = 120**

**Ha: mean < 4 Ha: mean != 4 Ha: mean > 4**

**Pr(T < t) = 0.0677 Pr(|T| > |t|) = 0.1353 Pr(T > t) = 0.9323**

**We are being asked to assess how likely it is that in the world as a whole, the regional average proportion of social media networks used is 4 (rather than the 3.86 of our sample). Given the sample of regions in our dataset, there’s only a 13.5% chance that we’d have gotten these observations if the true average were 4. That’s around a 1 in ~7.5 (1/0.135 = 7.4) chance.**

**More formally stated, again we fail to reject the null hypothesis that the true global cross-regional average is really 4. The p-value of 0.135 is well beyond the conventional alpha threshold of 0.05, and the value of 4 lies well within our 95% confidence interval range.**

**Which formulation is better? Again, I think saying that there is a 2 in 15 chance (i.e. 1 in ~7.5) that we’d have gotten these observation if the real cross-regional average were 4 is more intuitive and informative. Though technically speaking, both are correct.**

**TWO-SAMPLE T-TEST**

T-tests are more interesting when they are used to compare samples.

For example, if we are interested in the hypothesis that people are more likely to believe conspiracy theories in societies without a history of authoritarian rule, we could generate a variable for whether countries spent most of the last century as liberal democracies or otherwise.

**gen democratic\_legacy = 0**

**replace democratic\_legacy = 1 if country=="Australia" | country=="Denmark" | country=="France" | country=="Sweden" | country=="United States" | country=="United Kingdom" | country=="India" | country=="Japan" | country=="Germany"**

Now we can investigate this hypothesis formally by typing:

**ttest conspiracies, by(democratic\_legacy)**

The “by” option tells Stata that we are performing this test between the different categories of the variable “democratic\_legacy”. Check that you understand all the sections of this command, then look at the output. This time, the differences are expressed in terms of differences between the groups. The difference is statistically significant Pr(T > t) = 0.00

Open and run the do file, “Problem\_Set\_2\_dofile.do” that is in the Problem Set directory on Moodle (same folder as this document).

It will generate and code a variable called “developing,” which is a dummy variable, coded to 0 if a region has GDP per capita at PPP *above* $30,000, and 1 if that region has GDP per capita at PPP *below* that level. (Because this variable is coded at the regional level, many middle-income countries contain a mix of both “developed” and “developing” regions).

Add an “if” statement to perform this test for low development regions (variable: **developing==1**) countries and then for developed regions (**developing==0**). What do you find?

|  |
| --- |
| **Here we are running two t-tests, one on the sub-sample of regions with income per capita below $30,000 (developing == 1) and another on the sub-sample with income per capita above $30,000 (developing == 0).**  **Let’s look at them one by one.**  **ttest conspiracies, by(democratic\_legacy), if developing==0**  **Two-sample t test with equal variances**  **------------------------------------------------------------------------------**  **Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]**  **---------+--------------------------------------------------------------------**  **0 | 20 .1751848 .0086874 .0388513 .1570018 .1933678**  **1 | 39 .1441584 .0038676 .024153 .1363289 .1519879**  **---------+--------------------------------------------------------------------**  **combined | 59 .1546758 .0043102 .0331073 .146048 .1633036**  **---------+--------------------------------------------------------------------**  **diff | .0310263 .0082144 .0145774 .0474753**  **------------------------------------------------------------------------------**  **diff = mean(0) - mean(1) t = 3.7771**  **Ho: diff = 0 degrees of freedom = 57**  **Ha: diff < 0 Ha: diff != 0 Ha: diff > 0**  **Pr(T < t) = 0.9998 Pr(|T| > |t|) = 0.0004 Pr(T > t) = 0.0002**  **First, we are comparing regions in countries with a legacy of liberal democratic governance to those in regions without such a legacy, with respect to the mean (average) number of conspiracy theories. There is a difference in sample means – 0.144 in the first group, and 0.175 in the second. But does this reflect a difference in the general population, or could it simply be down to sampling?**  **The p-value of 0.0004 here tells us that, if there really were no difference between these two groups in the world as a whole, there would be only a 0.04% chance of getting the sample that we did. Alternatively stated, in comparable circumstances, we’d get such a result on only 1 in every 2500 occasions.**  **Note that this is with respect to the (null) hypothesis that there is no difference between the two groups in the general population. What about the hypothesis that “people are more likely to believe conspiracy theories in societies without a history of authoritarian rule”? This is one-tailed test. We would only be looking at whether the difference between the two groups is greater than or less than zero.**  **The p-value of 0.0002 for the (null) hypothesis that the difference is at less than zero tells us that, in a world in which people *were* more likely to believe conspiracy theories in societies without a history of authoritarian rule (i.e. in democracies), we’d get similar results in only 1 out of every 5000 samples. So that’s pretty strong evidence that we do not live in such a world. Among developed economy regions, we can reasonably conclude that believe people more conspiracy theories in non-democratic than in democratic areas.**  **ttest conspiracies, by(democratic\_legacy), if developing==1**  **Two-sample t test with equal variances**  ------------------------------------------------------------------------------  Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]  ---------+--------------------------------------------------------------------  0 | 51 .2078848 .0071473 .0510418 .1935291 .2222405  1 | 8 .1813681 .0068238 .0193006 .1652323 .1975038  ---------+--------------------------------------------------------------------  combined | 59 .2042893 .0063442 .048731 .19159 .2169887  ---------+--------------------------------------------------------------------  diff | .0265168 .01836 -.0102486 .0632821  ------------------------------------------------------------------------------  diff = mean(0) - mean(1) t = 1.4443  Ho: diff = 0 degrees of freedom = 57  Ha: diff < 0 Ha: diff != 0 Ha: diff > 0  Pr(T < t) = 0.9229 Pr(|T| > |t|) = 0.1541 Pr(T > t) = 0.0771  **Now for developing economy regions. Here the p-value of 0.0771 tells us that in a world in which people in democracies believed just as many or more conspiracy theories as in non-democracies, we would observe these results in just 7.71 out of every 100 samples (i.e. 7.7% of the time, or in about 1 in every 13 draws). That’s probably not enough to allow us to be sure that there really is a difference among developing economy regions in this regard (i.e. to reject the null hypothesis), even though, based on these results, we might consider it fairly likely.** |