This assignment is going to use **THREE different datasets**! Notice that the first two are individual-level datasets (surveys) and the second is country-level data (country characteristics and aggregated individual stats by country).

1. 1)  The first is the same as last time: your individual sample of the Census Microdata File.
2. 2)  For Questions 3–5, you will American National Election Study Data from 2016. It is posted on Canvas under /modules/course materials under course materials. *Trump, BLM, and Health Insurance attitudes! Hopefully that’s interesting right now!*
3. 3)  For Question 6, use the “Quality of Government” dataset. It is in the same place as the other data.

After writing the answers in a document you save for yourself, submit the answers in the appropriate question box on Canvas. Don’t just do answers in Canvas – save them!

**Q1: Proportion implied Frequency & Margin of Error using census data. [5pts]**

Estimate the number of Canadians who moved provinces over the last 5 years.

The data has a variable called “pr” for province of residence. It’s coded with the standard StatsCan coding, from 10 for NL to 60 for Yukon, NWT, and Nunavut. There’s another variable coded the same way called “pr5” which asks which province the person lived in five years ago.

You will need to generate a new dummy variable for this and use the not equal operator (!= or ~=) to classify people into the movers and stayers categories.

1a. (3 points) Use your census dataset sample to estimate the NUMBER (not the percentage) of people who live in a different province from 5 years ago. So: first find the proportion of people who live in a different province and then turn it into a number of people using the assumption that the total population of Canada is exactly 35 million. (You have no other source to help you estimate the number, just the proportion in your sample).

(2 points) 1b. Indicate how far away from the true number of people who moved provinces in the full population (i.e. not the percentage) that you would expect to be, 19 times out of 20.  
Say: ± \_\_\_\_\_\_ number of people. (NOT ± %) You can calculate this using the formula for the standard error of a proportion and then use that result to calculate the number of Canadians, as you will have done in question 1a.

In Canvas, just enter one number for a) and another number for b). For this question you do not need to ‘write up’ the answers.

**Q2. Means from the census [5 pts]**

2a. (3 points) Compare the Government Transfers income (gtrfs) for people whose first language was English and whose first language was French (exclude the Both and Neither categories on that variable). Report the two means and the difference between them a smoothly worded paragraph that summarizes the findings for a reader.

2b. (2 points) Add a sentence reporting the standard error of the mean government transfer income for those whose first official language spoken was just English. Your report of this number should be in a sentence that says how close you expect to be to the population mean, in 95% of samples. Calculate this with the formula, using the mean, the standard deviation, and the number of cases (n) in the English category of that variable.

**Questions 3 to 5: American National Election Study.**

Now switch to the 2016 American National Election Study (ANES 2016) dataset.

DO THIS FIRST:  
Draw a random sample of 2,500 cases from the dataset. You will all get different

samples that we can replicate to check your answers because of the way random number generation works in computers.

First: set the random number seed by typing: set seed *courseidnumber* (where you replace ‘courseidnumber’ with the same number as your course id (The number between 1 and 195, *NOT your real student number*).  
Then: use the command *sample*: sample 2500, count

Now use the command called count to double check that you now have 2500 cases to work with: just type count and hit enter  
Stata should simply report a number close to 2500. (If it’s close, it’s ok).

**Q3. Americans’ interest in politics and feelings about groups [5 points]**

We are going to look at the relationship between Americans’ voting intention (before the 2016 election) and their attitudes toward the Police and BLM. To start, use the lookfor command to find the two variables that measure survey respondents’

evaluation of the POLICE and BLACK LIVES MATTER 0-100 thermometer scale. Both of those variables have missing values that are strangely coded, so you’ll need to use this command to set all values outside 0 to 100 to missing. (You put in the variable names – actually numbers – where I have “*varname*”).

replace *varname* =. if *varname*<0 | *varname*>100

Now that you have those variables only from 0 to 100 (check with summarize), create a new variable that indicates how much more positive the person is toward BLM compared to their feeling about the Police. I’ll call this BLMvsP for short here.

Next, find the variable that measures “For whom does R intend to vote for President” [just type lookfor intend and find it in the list that comes up]. Create a new variable called something like “voteint” that has only 3 categories: Clinton, Trump, and all other responses.

In a single, concise, and engaging paragraph, report the mean score for BLMvsP variable for each of the three categories of the vote intention variable. Your audience is someone reading a newspaper article or op-ed. You’ll need to say how the BLMvsP variable measures what it measures, referring to the range of possible values it could be (i.e the extremes and neutral point). Give your Class ID number in parentheses at the start or end of your answer.

NOTE:  
1. Look at the results for the mean BLMvP score for Clinton very carefully. Don’t get fooled by Stata omitting the leading 0 for a value between -1 and 1. (Some of you may get a result between -1 and 1.  
2. If you want to check to reassure yourself you did calculate the BLM-positive-vs- POLICE variable correctly, use the tabulate command with the summarize option to see the means of your BLMvP variable broken down by the variable that asks people if they self-identify as Black or African-American (V161310b).

**Q4: Partisanship, perceptions of politicians, and crosstabs [5 points]**

Next let’s look at the relationship in the US 2016 Data (ANES 2016) between thinking the country is on the “right track” [type lookfor track to find it] and having health insurance [lookfor insurance]. Without creating a new variable, just recode these so that the Refused and Don’t Know categories are missing.

Run a crosstab that shows whether having health insurance and opinion about the country being on the right track were (or weren’t) related in 2016 in the US. Report results of this crosstab in a clear and compelling short paragraph where you briefly say what your variables are. You do not need to run a chi-square test, or report the p-value, for this question.

Give your Class ID number in parentheses at the start or end of your answer.

**Q5: Interpreting p-values [5 pts]**

Rerun the cross tab you did for Q4 but separately only among Democrat indentifiers and then only among Republicans. You can use the if statement in the tab command just like you’ve done before. The Democrats are V161155==1. Republicans are V161155==2.

When you run the crosstabs these two times use the , chi option at the end of the commend to get a p-value (from a *chi-square test*). The p-value you want is the one p- value from the middle hypothesis test at the bottom [Ha: diff!=0]1. In a nicely sewn- together paragraph, present the two sets of results. What is the relationship in the crosstab for Democrats and what is it for Republicans. After discussing the relationship, then for each of them report and interpret the p-value for an intelligent reader who knows little about statistics.

Give your Class ID number in parentheses at the start or end of your answer.

**Q6: Difference of means [5 pts]**

Use the Quality of Government data set (with COVID data).

First, drop a few countries. Use the Stata drop in command to drop cases from the data. Start with your class ID number and add 10 to it so you’re dropping 11 countries. For example, if you class ID number were 87 you’d type:

1 That’s shorthand for: *The alternative hypothesis (Ha) is that the difference of the means in the two groups is not equal to zero (!=0).*

drop in 87/97

and Stata will say (11 observations deleted)

Now:

1. run a t-test to assess the difference in per capita COVID cases among countries

with and without a proportional representation electoral system. Use totcases\_pc to measure cases (up to date as of Oct 18) and dpi\_pr for the proportional system variable (proportional = 1). Report results from the t-test in a smoothly worded paragraph that explains your findings. Be sure to provide the readers with sufficient information (imagine they know nothing about the data set or variable) to understand your results, including the two means turned into something easier to understand than just the mean number, which is a proportion (zero to 1). Talk about whether or not there is a relationship, and its size and direction. Also, be sure to interpret the p-value from the middle hypothesis test [Ha: diff!=0].

1. Then briefly do the same thing but for the total deaths per capita variable: totdeaths\_pc.