

Assignment 1

Due Date: Thursday, October 13th (Submitted through MyCourses by Midnight ET)

This assignment is worth 15% of your final grade

The assignment needs to be turned in, in class at the start of class October 13^h, unless you meet an exception criterion that has already been discussed with the instructor, no late assignment is accepted. If you have a scheduling conflict on the day when the assignment is due, make sure to submit your assignment beforehand.

You are expected to submit a write-up of your results, which includes **two** different documents:

- a) The answers to the questions in a **Word document** (or PDF, **NO** Pages files pc users can't open these). The answers include the values you obtained, the tables or graphs as required, and the interpretation of your results.

In the case of tables, you can cut and paste the table you obtained from Stata. However, if you choose to do so, make sure the formatting of the table does not change as you cut and paste from Stata to Word. To preserve a good Stata layout after you paste the Table in Word, you can choose the font "Courier New" and pick a font size of 9 or 10 depending on how big the table is. Sometimes, it might be easier to create a new table in Word and input the values obtained from Stata in the cells – this is your choice, but keep in mind that the information needs to be presented in a clear and organized manner.

In the case of graphs, you should also be able to copy and paste them.

It is your responsibility to ensure these files work and can be opened by myself and the TAs.

- b) Your **do-file** with the syntax you used. Organize your do-file in a clean manner, and feel free to add comments and notes to it.

For all written answers, round values to two decimal places (e.g. a number such as 2.1654 can be rounded up to 2.17)

You can discuss the assignment with other students in the class, but you must write it up independently. Both the write-up and the do-file need to be written independently. Any deviation from this constitutes cheating.

Imagine you are a researcher interested in learning more about the condition of Canadian workers in Canada in 2015 – their income, labor force participation status, hours worked, etc. For this, you employ Statistics Canada’s Canadian Income Survey (CIS), 2015.

Use Stata and the Canadian Income Survey (CIS) 2015 dataset posted on MyCourses (CIS_2015_modified.dta) to answer the below questions. Refer to the codebook for this dataset (posted on MyCourses, codebook_modified_CIS_2015.pdf) for additional information on the variables.

Question 1 [5 pts]

Load the dataset in Stata. How many individuals are there in the dataset? How many variables?

Question 2 [10 pts]

Focus on the variable PROV, which stands for province. Which kind of variable is this? Use Stata to produce a frequency distribution table for this variable and copy your results in a text editor. Which is the province where the majority of individuals live, report it and its percentage? What proportion of individuals live in Saskatchewan?

Question 3 [5 pts]

Plot a graph of the PROV variable and – if you deem it appropriate – order the bars accordingly. Report the graph in the word file and the syntax you used to produce it.

Question 4 [15 pts]

Consider now the MTINC variable, which indicates annual market income. Which kind of variable is this? Produce a graph (in Stata) that is appropriate for this type of variable. Explain the reasoning behind your choice. Provide the syntax you used to produce the graph, copy and paste it below your text. Interpret your graph.

Question 5 [10 pts]

Use Stata to calculate two measures of central tendency for the MTINC variable. Specify the syntax you used. List the values you obtained, compare them, and interpret in words what the comparison between the two measures tells you about the distribution of your data.

Question 6 [15 pts]

Now, focus on measures of dispersion. Using Stata, calculate both the standard deviation and the 5-number summary (5NS) for the MTINC variable. Specify the syntax you used and report the values you obtained. Based on the shape of the distribution (which you identified in Question 5), which measure of variability do you think is more appropriate for the MTINC variable? Why? Explain your reasoning.

Question 7 [5 pts]

Does the average annual market income (MTINC) differ for men and women in the dataset (the sex of the person is recorded in the variable SEX)? If so, report the two averages (one for men and one for women) and interpret them in words. For instance, discuss whether any observed difference is in line with what you would expect and why.

Question 8 [15 pts]

Now let's focus on the variable USHRWK, which gives the average hours worked per week during the year 2015. Use Stata to draw a boxplot of this variable and interpret what you see. Report the graph in your text document. (Note: If the median line does not show up, it's because the median coincides with either Q1 or Q3).

Help: To better visualize the data and make your life easier, use the `nooutsides` option in Stata.

Question 9 [10 pts]

Do the boxplots differ by sex of the individual (variable SEX in the dataset) and by immigrant status (variable IMMSTP in the dataset)? Use Stata to draw these boxplots, report them in your text document, and interpret what you observe.

Help: To better visualize the data and make your life easier, use again the `nooutsides` option

Question 10 [5 pts]

Is there any other variable in the dataset – one that we haven't mentioned up until now – for which it would make sense to calculate a mean and Standard deviation? The name of the variable and a motivation behind your choice will suffice as an answer. No need to actually compute the summary measures.

Question 11 [5 pts]

Identify another variable in the dataset – one that we haven't mentioned up until now – for which it would make sense to draw a bar graph. The name of the variable with a motivation is enough here. No need to actually do the graph.