**LPO 7870 Research Design and Data Analysis II, 2024**

**Assignment 1**

**Submission Guidelines:**

* **Please submit an electronic copy of your group’s solutions as a Microsoft Word (MS) document file by February 5, at 4:00pm**
* **All submissions via Brightspace**
* This assignment is done by pairs. These pairs were created by random allocation. If you prefer to work alone, please let the teaching team know.
* You should submit one response, but you need to put the two names.
* Please submit your answers in Brightspace, “Assignments” section. Word or PDF are fine. You can copy and paste the main Stata outcomes.
* Pro-tip: if you use Courier New font, Stata output will line up as it does in the output window.

**QUESTION 1.**

A standardized test is given to 453 randomly selected 6th grade students in Florida. The mean score on this test is 1,013 with a standard deviation of 108.

1. Construct a 95% confidence interval for the population mean test score of 6th grade students in Florida.
2. Without doing any calculations, explain whether (and why) a 90% confidence interval for the population mean would be wider or narrower than the 95% confidence interval.
3. It is decided that a sample of 1,000 students instead of 453 students will be taken. Without doing any calculations explain whether (and why) a 95% confidence interval calculated using 1,000 students would be wider or narrower than the 95% confidence interval calculated using 453 students.

Suppose now that another group of 503 6th grade students are selected at random in Florida. These students, however, are given a 3-hour preparation course before the test is administered. The mean score for this group is 1,019 with a standard deviation of 95.

1. Construct a 95% confidence interval for the population mean test score associated with the prep course.
2. Test the hypothesis that there is significant difference between the average test scores of students that received the prep course and those that did not receive the prep course. Use a 95% confidence level.

**QUESTION 2**

The file cps\_2008.dta contains data on the earnings and demographic characteristics of 10,651 working adults in 2008. A description of the variables in the dataset can be viewed by using the describe command in STATA. The data was compiled from research files available for download at the Current Population Survey (CPS) website: <http://www.census.gov/cps/>.

Download the dataset and load the data into STATA. To get a quick look at all the variables in the data, type “describe” to see a description of each variable name, and any descriptive label it may have been given. Familiarize yourself with the variables in this dataset using commands like “summarize” (for quantitative variables) and “tabulate” (for categorical variables).

1. The dataset contains a variable called earnings\_pw which is the total earnings of an individual in a given week. However, hourly wages are often thought to be a better measure of economic status because wage comparisons aren't affected by differences in the amount of hours worked over a year. Construct a new variable called wages by typing “generate wages=earnings\_pw/hours\_pw”. The new variable wages now measures the hourly wage rate for individuals in the dataset. What is the mean, min and max of hourly wages in the sample? What is the standard deviation of hourly wages?
2. Let’s examine the gap in hourly wages between men and women. Begin by calculating the average wage of men and women in the sample.
   * What is the difference in the sample mean wages between men and women?
   * Now conduct a t-test to test the null hypothesis that the average wages of men and women are equal. To do so using the command: ttest wage, by(sex). Report the results of this t-test and discuss your conclusion from this hypothesis test.
   * Now construct a 95% confidence interval separately for the mean hourly wages of men and women. Do these confidence intervals overlap?
3. Produce a scatterplot depicting the relationship between hourly wages and years of education. You can produce a scatterplot by typing: scatter wage educ, title(“Scatterplot of wages and years of education”). Do this and comment on the relationship between wages and years of education. Save your graph as a “PNG” or “TIFF” file and incorporate it into your problem set. Please include a brief paragraph discussing the graph.
4. We can summarize this bivariate relationship further if we compute the average wage at each year of educational attainment and plot this average by years of education. To do so, use the following commands: (1) sort educ (2) by educ: egen mwage = mean(wage) (3) scatter mwage educ, title (“Scatterplot of Average Wages by Years of Education"). Note that egen is a special version of the generate command that allows you to calculate group averages and other types of statistics. In this case, the egen command calculates the average wage by years of educational attainment. Once again comment on the relationship between years of education and average earnings. Also, once again save your graph as a “PNG” or “TIFF” file and incorporate it into your problem set. Please include a brief paragraph discussing the graph.
5. Now let’s use a simple (one variable) regression framework to assess the relationship between wages and education. First fit a simple regression model of wages on education. Interpret the coefficient and the t-stat on education.
6. Please discuss whether (and why) you agree or disagree with the following statement: “Given that the coefficient is positive and significant, an increase in education causes higher wages”.