**LPO 7870 Research Design and Data Analysis II, 2025**

**Assignment 1**

**Submission Guidelines:**

* **Please submit via Brightspace an electronic copy of your solutions as a Microsoft Word or PDF document by January 27, at 4:00pm.**
* This assignment may be completed individually or in pairs.
* If you work in pairs, you should submit one response, but you need to put both names on the document.
* You should copy and paste the main Stata outcomes such as graphs and tables. Pro-tip: if you use Courier New font, Stata output will line up as it does in the output window.
* You must also submit a .do file that provides the Stata code used to complete the assignment and answer each data analysis question. The .do file should have comments that identify which part of the code answers which question.

**Question 1: Inference about population means** (15 points)

Random samples of 50 self-identifying male and 50 self-identifying female students were administered a survey instrument to assess their self-concept in 8th grade. The male students scored an average of 23.2 on the assessment (with a standard error of 0.97) while the female students scored an average of 19.3 (with a standard error of 1.16).

1. Provide an interpretation (in words) of the standard errors reported above. (3 points)
2. Based on these samples, construct 95% confidence intervals for the *population* mean self-concept for male and female 8th graders, respectively. Interpret. (5 points)
3. 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 found in part b. (3 points)
4. Your colleague asks whether the estimated difference in mean self-concept between male and female students is *statistically significant* (at the 95% level)*.* What would you answer and how did you determine this? (4 points)

**Question 2: Minimum sample size** (5 points)

You are preparing to collect survey data to estimate families’ mean travel time from home to school in a large urban school district. You would like your estimate to be correct to within 2 minutes in 95% of random samples. How large of a sample is needed to meet this requirement? Hint: A previous study of commuting time with a similar population suggests that 15 minutes is a reasonable approximation for the standard deviation (σ) of travel time to school. Show or explain your reasoning.

**[OVER]**

**Question 3: Stata exercise – descriptive and inferential statistics** (24 points – 4 each)

The file 25\_cps\_2008.dta contains data on the earnings and demographic characteristics of 10,651 working adults in 2008. A description of the variables and any labels the variables have been given 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](https://www.census.gov/programs-surveys/cps.html).

Download the dataset and load the data into Stata. Familiarize yourself with the variables in this dataset including what they measure, their level of measurement, their value labels, and any missing data.

1. Produce a histogram of the variable *hours\_pw*. What do you notice? Are there nonsensical values? If so, you should modify the variable to account for those values and reproduce the histogram. Explain what you did and the implications of that action. (4 points)
2. 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 number of hours worked per week. Construct a new variable called *wages* by dividing *earnings\_pw* by *hours\_pw*. The new variable wages now measures the hourly wage rate for individuals in the dataset. How many observations does this new variable have? Is that different from the total number of observations in the dataset? If so, why are they not the same? (4 points)
3. What is the mean, min and max of hourly wages in the sample? What is the standard deviation of hourly wages? What is the median of hourly wages? (3 points)
4. 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 hourly wages between men and women?
   * Conduct a t-test to test the null hypothesis that the average wages of men and women are equal. Report the results of this t-test and discuss your conclusion.
   * Now construct a 95% confidence interval separately for the mean hourly wages of men and women. Do these confidence intervals overlap?
5. Produce a well-labelled scatterplot depicting the relationship between hourly 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 solutions. Please include a brief paragraph discussing the graph.
6. 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 and retain these as part of your dataset. 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.