**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 need3/ 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. **(3 points)**

We are 95% confident that the mean test score for all 6th graders in Florida is between 1,023 and 1,003. “95% confidence” comes from the fact that—in 95 percent of random samples—a confidence interval constructed using this formula will contain the true population mean.

1. 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.

**(6 points)**

A 90% confidence interval would be **narrower** than a 95% confidence interval since the t value for a 90% confidence interval (1.64) is smaller than the t value for the 95% confidence interval (1.96).

Conceptually: Sample means are random variables, meaning that if were to take multiple samples from the same population we would calculate different sample means. We construct a 95% confidence interval so that the intervals constructed from 95% of samples will include the true population mean. 90% confidence intervals are narrower, as they only contain the true population mean in 90% of samples (a narrower range of values will accomplish this).

1. 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. **(6 points)**

A 95% confidence interval calculated using 1,000 students would be **narrower** than a 95% confidence interval calculated using 453 students.

Since the standard error of is (, increasing the sample size (n) reduces the standard error and hence the width of the confidence interval. In other words, the larger the sample, the less the sample mean will vary around the true population mean.

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. **(5 points)**

We are 95% confident that the mean test score for students in Florida when they receive the test preparation course is between 1,011 and 1,027.

1. 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. **(5 points)**





This test statistic is approximately normal (with large enough sample sizes).

Since , we fail to reject the null hypothesis at the 5% level (α=0.05). Thus, we find no evidence that the prep course increased student performance. Note it is also possible to test for differences between two population means by comparing their confidence intervals. If two 95% intervals do not overlap, we can reject the null hypothesis of no difference at the 5% level. If they do overlap, we cannot reject the null.

**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 points)**

. summarize wages

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

wages | 10651 21.18958 12.36179 6.5 96.16666

1. 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? **(5 points)**

bysort sex: sum wages

-----------------------------------------------------------------------------------------------------

-> sex = Male

Variable | Obs Mean Std. dev. Min Max

-------------+---------------------------------------------------------

wages | 5,637 22.99682 13.21108 6.5 82.42857

-----------------------------------------------------------------------------------------------------

-> sex = Female

Variable | Obs Mean Std. dev. Min Max

-------------+---------------------------------------------------------

wages | 5,014 19.15779 10.98292 6.5 96.16666

. ttest wage, by(sex)

Two-sample t test with equal variances

------------------------------------------------------------------------------

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

---------+--------------------------------------------------------------------

Male | 5637 22.99682 .1759602 13.21108 22.65187 23.34177

Female | 5014 19.15779 .155105 10.98292 18.85371 19.46186

---------+--------------------------------------------------------------------

combined | 10651 21.18958 .1197805 12.36179 20.95479 21.42437

---------+--------------------------------------------------------------------

diff | 3.83903 .2370821 3.374305 4.303755

------------------------------------------------------------------------------

diff = mean(Male) - mean(Female) t = 16.1928

Ho: diff = 0 degrees of freedom = 10649

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 1.0000 Pr(|T| > |t|) = 0.0000 Pr(T > t) = 0.0000

Note that the t for the null hypothesis that the population mean of wages for men and women is the same is 16.1928. Since we have over 10,000 observations, the critical value for the hypothesis test is 1.96. Thus we firmly reject the null that the population mean wage of men and women is the same.

. ci means wages if sex == 1

Variable | Obs Mean Std. err. [95% conf. interval]

-------------+---------------------------------------------------------------

wages | 5,637 22.99682 .1759602 22.65187 23.34177

. ci means wages if sex == 2

Variable | Obs Mean Std. err. [95% conf. interval]

-------------+---------------------------------------------------------------

wages | 5,014 19.15779 .155105 18.85371 19.46186

1. 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 points)**



Years of education is in the range of 0 to 21. There are two “pick” densities: at 12 (high school) and at 16 (college). Wages seem positively associated with years of education, where high wages are especially found among those with high education levels. However, the minimum wage for each level of education is relatively consistent.

1. 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. **(4 points)**



By associating years of education with average wage, the relationship becomes more clear. There is a positive association between years of education and average wage. This relationship may be linear or mildly curvilinear in that falling below the threshold of high school graduation flattens the positive distribution of wages.

1. 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. **(5 points)**

. reg wages educ

Source | SS df MS Number of obs = 10,651

-------------+---------------------------------- F(1, 10649) = 2659.78

Model | 325251.353 1 325251.353 Prob > F = 0.0000

Residual | 1302215.71 10,649 122.285258 R-squared = 0.1999

-------------+---------------------------------- Adj R-squared = 0.1998

Total | 1627467.07 10,650 152.813809 Root MSE = 11.058

------------------------------------------------------------------------------

wages | Coefficient Std. err. t P>|t| [95% conf. interval]

-------------+----------------------------------------------------------------

educ | 2.130196 .0413045 51.57 0.000 2.049231 2.21116

\_cons | -8.206163 .579967 -14.15 0.000 -9.343007 -7.069319

------------------------------------------------------------------------------

For every one unit increase in years of education, there is a 2.13 dollar increase in wages (earnings per hour). The standardized test-statistic for education is 51.57, which is much greater than 1.96 and therefore we can conclude that increasing one’s years of education has a statistically significant impact on wages.

1. 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”. **(5 points)**

I disagree with this statement due to the fact that we have not accounted for potentially confounding factors. In this way, education could be associated with wages through some other measurable variable. One example of such a variable is gender. We saw above that wages vary systematically by gender. If a variable like this is not appropriately accounted for, we could be under- or over-estimating the impact of education on wages.

Other issues include that causality implies a directional relationship where one’s years of education is decided before they receive these wages. In certain scenarios, like a graduate TA, these could be concurrent changes.