Who is your TA (check one)?

🞏 De Andre 🞏 Corey 🞏 Jingkai 🞏 Neeraj

**SOC 3811/5811 – STATA Assignment #2**

PART 1. ONLY FOR STUDENTS IN SOC **5811**

Go the GSS Data Explorer website at: <https://gssdataexplorer.norc.org/>. You may need to register for a (free) account. Once you are logged in, go to “MyGSS” and create a new project. When prompted, specify that you want “Stata (Control + System)” files (and not files for other software packages). Your cart will automatically include three variables, including “year.” Add the variables ADHDREAL and AGEWED to your cart. **Before going any further, be sure to read the on-line documentation well enough to understand what survey questions were asked to generate these two variables and to ascertain value labels and missing data codes**. Then, create a data extract that includes the variables ADHDREAL and AGEWED.

Once the extract is ready, the GSS Data Explorer will provide Stata code for reading in the two variables. Edit as necessary, and be sure to declare missing values to be missing!Then do the following:

1. Have STATA report a 95% confidence interval for the population proportion of people who think that ADHD is a real disease



1. Have STATA report a 99% confidence interval for the population mean age at first marriage. (Paste all of your syntax below)



1. Interpret—in words—those two confidence intervals

For ADHD, we are 95% certain that the population proportion of people who think that ADHD is a real disease is between 0.760 and 0.816. For age at first marriage, we are 99% certain that the mean age at first marriage is between 22.02 and 22.17.

1. Paste below the Stata syntax you used to read in the GSS extract and perform the analyses above.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands tell STATA how to interpret things

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

clear

cd "C:\Users\warre046\Dropbox\1. TO DO !\\_Fall Stats Class\3 Assignments"

set more off

log using "STATA Assignment 2 for 5811.log", replace

log on

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* These commands read the raw data and label things

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

infix agewed 9-10 adhdreal 11-11 using "GSS Extract.dat"

label variable agewed "Age When First Married"

label variable adhdreal "Is ADHD a Real Disease?"

label define ar 1 yes 2 no

label values adhdreal ar

replace adhdreal=. if adhdreal==8 | adhdreal==9

replace agewed=. if agewed==98 | agewed==99

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* THIS SYNTAX RECODED ADHDREAL TO EQUAL 1 FOR PEOPLE WHO THINK ADHD IS REAL, 0 OTHERWISE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

replace adhdreal=0 if adhdreal==2

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* THIS SYNTAX REPORTS A 95% CI FOR THE PROPORTION OF PEOPLE WHO THINK ADHD IS REAL

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ci proportion adhdreal

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\* THIS SYNTAX REPORTS A 99% CI FOR THE MEAN AGE AT FIRST MARRIAGE

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

ci means agewed, level(99)

log off

log close

PART 2. FOR ALL STUDENTS IN BOTH SOC **3811** AND **5811**

In this exercise, you will use data from the American Community Survey (ACS). The ACS is a product of the U.S. Census Bureau, and involved interviewing millions of Americans each year. For an introduction to the ACS, go to the ACS website ([here](https://www.census.gov/programs-surveys/acs)).

For this exercise, I have constructed a data file that includes two variables collected from everyone who responded to the ACS in 2010 and who lived in one of two metropolitan areas: Minneapolis/St Paul and Duluth/Superior. The two variables are: (1) People’s poverty status and (2) How long it takes people to commute to work.

Use STATA syntax files that you already have (from the first assignment, or from class examples) and modify them to accomplish the following goals. When you are done, type or paste your answers for the questions below into a word processor (e.g., Microsoft Word) and turn in the assignment via Canvas.

1. Read the data file for this assignment into STATA

Done

1. Be sure to declare “zero” to be a missing value for TRANTIME, the time it takes people to get to work

Done

1. Create a new dichotomous poverty variable that equals “1” if the ratio of people’s income to the poverty line (POVRATIO) is less than 100 and “0” otherwise; see the bottom of the assignment for an example of how to do this in STATA

Done. See STATA code.

1. Separately for Minneapolis/St Paul and Duluth/Superior, produce (1) a histogram of the “time to work” variable; (2) measures of central tendency and spread for the “time to work” variable; and (3) a frequency distribution for the “poverty status” (0 vs 1) variable

**DULUTH: MPLS/ST PAUL:**

 

*Time to Work*

Mean: 22.5 SD: 22.5 Mean: 24.9 SD: 18.9

*Poverty Status*

1=Poor: 14.6% 1=Poor: 11.1%

0=Not Poor: 85.4% 0=Not Poor: 88.9%

1. Separately for Minneapolis/St Paul and Duluth/Superior, use STATA code to produce (1) a 95% confidence interval for the mean of “time to work” and (2) a 95% confidence interval for the proportion who are poor; ; see below for an example of how to do this in STATA





Use the results from #4 above to do the following:

1. Separately for Minneapolis/St Paul and Duluth/Superior, produce by hand (1) a 95% confidence interval for the mean of “time to work” and (2) a 95% confidence interval for the proportion who are poor

**FOR TRANSPORTATION TIME:**

FOR DULUTH: $\overline{Y}\pm t\_{a/2}\frac{s\_{Y}}{\sqrt{N}}$= $22.5\pm 1.984\frac{22.5}{\sqrt{901}}$= 22.50±1.49.

Thus, we can say with 95% certainty that the population mean in Duluth is between 21.01 and 23.99. These differ from STATA’s answer only because of rounding.

FOR MPLS/ST PAUL: $\overline{Y}\pm t\_{a/2}\frac{s\_{Y}}{\sqrt{N}}$= $24.9\pm 1.962\frac{18.9}{\sqrt{10,228}}$= 24.90±0.37.

Thus, we can say with 95% certainty that the population mean in MPLS/St Paul is between 24.53 and 25.27. These differ from STATA’s answer only because of rounding.

**FOR POVERTY STATUS**:

FOR DULUTH: $\hat{p}$ equals 0.146.

So, $\hat{p} \pm Z\_{α/2} × \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} $🡺$0.146 \pm 1.96 × \sqrt{\frac{0.146(1-0146)}{2127}} $🡺0.146±0.015.

Thus, we can say with 95% certainty that the population proportion is between 0.131 and 0.161. This differs just slightly from STATA’s answer because of rounding.

FOR MPLS/ST PAUL: $\hat{p}$ equals 0.111.

So, $\hat{p} \pm Z\_{α/2} × \sqrt{\frac{\hat{p}(1-\hat{p})}{N}} $🡺$0.111 \pm 1.96 × \sqrt{\frac{0.111(1-111)}{21,352}} $🡺0.111±0.004.

Thus, we can say with 95% certainty that the population proportion is between 0.107 and 0.115. This differs just slightly from STATA’s answer because of rounding.

1. Confirm that your answers to #5 and #6 are the same

Except for slight rounding errors, they are the same

1. How do you interpret the confidence intervals in #5/#6 above?

**FOR TRANSPORTATION TIME:**

FOR DULUTH: We can say with 95% certainty that the population mean amount of time people spend commuting in Duluth is between 21.01 and 23.99 minutes.

FOR MPLS/ST PAUL: We can say with 95% certainty that the population mean amount of time people spend commuting in MPLS/St Paul is between 24.53 and 25.27 minutes.

**FOR POVERTY STATUS**:

FOR DULUTH: We can say with 95% certainty that the population proportion of people in poverty in Duluth is between 0.131 and 0.161.

FOR MPLS/ST PAUL: We can say with 95% certainty that the population proportion of people in poverty in MPLS/St Paul is between 0.107 and 0.115.

**DESCRIPTION OF VARIABLES IN “STATA ASSIGNMENT 2.DAT”**

**METAREAD** (Column 4-7)

Metropolitan Area

2240=Duluth-Superior, MN/WI

5120=Minneapolis-St. Paul, MN

**POVRATIO** (Column 18-20)

Ratio of Person’s Income to the Poverty Threshold

 <100 = Below Poverty Line

 100 = At Poverty Line

 >100 = Above Poverty Line

**TRANTIME** (Column 21-23)

Travel time to work

 0 = Zero minutes

 1= 1 Minute

 Etc.

**HOW TO CREATE A NEW VARIABLE IN STATA**

Imagine that X is a continuous variable that ranges from 0 to 10. Also imagine that you want to create a dichotomous variable … Z … that equals “0” if X is between 0 and 5 and “1” if X is between 6 and 10. How would you do that? Below I provide STATA syntax that would work; the STATA syntax assumes you use a semi-colon after commands. You will need to modify this code to suit your purposes. Note that in the first lines of the code I create a new variable; this may be useful to you later. Also, in both cases, X will be missing if Z is missing.

gen X=. ;

replace X=0 if Z>=0 & Z<=5 ;

replace X=1 if Z>=6 & Z<=10;

**HOW TO CREATE CONFIDENCE INTERVALS IN STATA**

Imagine that Y is a continuous variable and that X and Z are dichotomous variables. The code below produces 95% confidence intervals for Y and Z separately for each value of X (which can equal 0 or 1).

ci means Y if X==0, level(95)

ci means Y if X==1, level(95)

ci proportions Z if X==0, level(95)

ci proportions Z if X==1, level(95)

If you want to produce confidence intervals with some level of confidence besides 95%, just replace the 95 in “level(95)” with your desired confidence level