Dis 7: Stata Q&A

Check out the solution from Dis 3 (Stata Review) when doing your Stata problem set.
You can use that discussion's solution and this as a guide when completing your Stata problem set.

1 Stata Exercise

Before we start the Q&A part of this section, let's look at one Stata exercise together. This exercise uses the same dataset given in your Stata Assignment (that's due Dec 10 @ 11pm on Canvas).

You are given a dataset that is created by a collaboration between the Robert Wood Johnson Foundation and the University of Wisconsin Population Health Institute.

The dataset contains variables from the 2019 Wisconsin County Health Rankings.

1. First thing first, let's make sure we can keep track of your work and your results.

   (a) To keep track of your results, clear out the Results panel.
   To do so, right click anywhere within the Results panel, and select "Clear results".

   (b) To keep track of your work / code, create a Do-file with the first line as a comment, and put your name down there.
   The easiest way to create a Do-file is by typing `doedit` in the Command panel, and then hit the enter / return key on your keyboard. This should open up a separate window that looks like a code editor. This is the Do-file editor.
   To comment out any part of the code, use the `*` symbol. This means that the first line in your Do-file editor should be

   ```
   * FirstName LastName
   ```

   (c) Set up the Do-file environment before proceeding.
   Some commands, like clearing Stata's memory, should be included at the beginning of every single Do-file. Personally, whenever setting up Stata's environment, I prefer to both
   - clear out Stata's memory (`clear all`), and
   - tell Stata to always report the full output for every single line of command (`set more off`)

   This means that the following lines should be added to your Do-file:

   ```
   * Set environment
   clear all
   set more off
   ```

2. The dataset has been given in `.csv` and `.dta`. The files are on Canvas under the following names:

   Econ 310 Stata Assignment Data.csv

   Econ 310 Stata Assignment Data.dta

   Load the data into Stata environment.
   Remember from our Stata Review section (Dis 3) that you always need to change your working directory before loading any dataset. This is to tell Stata where your dataset is located at. To change

your working directory, go to Stata's menu bar, select **File** > **Change working directory...**. It should open up a file selection dialogue. Navigate to the file folder containing the dataset, and click **Choose**.

Now we can load the dataset. Your actual Stata Assignment gives you directions on how to import data through the **File** menu. Here, let's import data using commands instead.

Recall from Dis 3 that if you want to load `.dta` files, then `use` command should be used. Otherwise, `import` command is appropriate. Thus, if you want to load "Econ 310 Stata Assignment Data.csv", then the command to type in Do-file should be

```
import delimited "Econ 310 Stata Assignment Data.csv", clear
```

If you want to load "Econ 310 Stata Assignment Data.dta" instead, then the command to type in Do-file should be

```
use "Econ 310 Stata Assignment Data.dta", clear
```

Remember to run the Do-file after typing down the commands to see their effects.

3. Describe your data using the **describe** command.

   The command to use is

   ```
   describe
   ```

   The result from running this command is the following:

```
. describe

Contains data
 Observations:            72
    Variables:            19
```

Variable name	Storage type	Display format	Value label	Variable label
fips	long	%12.0g		**FIPS**
state	str9	%9s		**State**
county	str11	%11s		**County**
physicallyunh~s	float	%9.0g		**Physically Unhealthy Days**
percentadults~s	byte	%8.0g		**Percent Adult Smokers**
percentphysic~e	byte	%8.0g		**Percent Physically Inactive**
percentexcess~g	byte	%8.0g		**Percent Excessive Drinking**
percentalcoho~g	byte	%8.0g		**Percent Alcohol-Impaired driving**
percentuninsu~d	byte	%8.0g		**Percent Uninsured**
percentfluvac~d	byte	%8.0g		**Percent Flu Vaccinated**
hsgraduationr~e	byte	%8.0g		**HS Graduation Rate**
percentsomeco~e	byte	%8.0g		**Percent Some College**
numberofunemp~d	int	%8.0g		**Number of Unemployed**
laborforce	long	%12.0g		**Labor Force**
percentunempl~d	float	%9.0g		**Percent Unemployed**
percentchildr~y	byte	%8.0g		**Percent Children in Poverty**
percentile~80th	long	%12.0g		**Percentile Income 80th**
percentile~20th	long	%12.0g		**Percentile Income 20th**
medianhouseho~e	long	%12.0g		**Median Household Income**

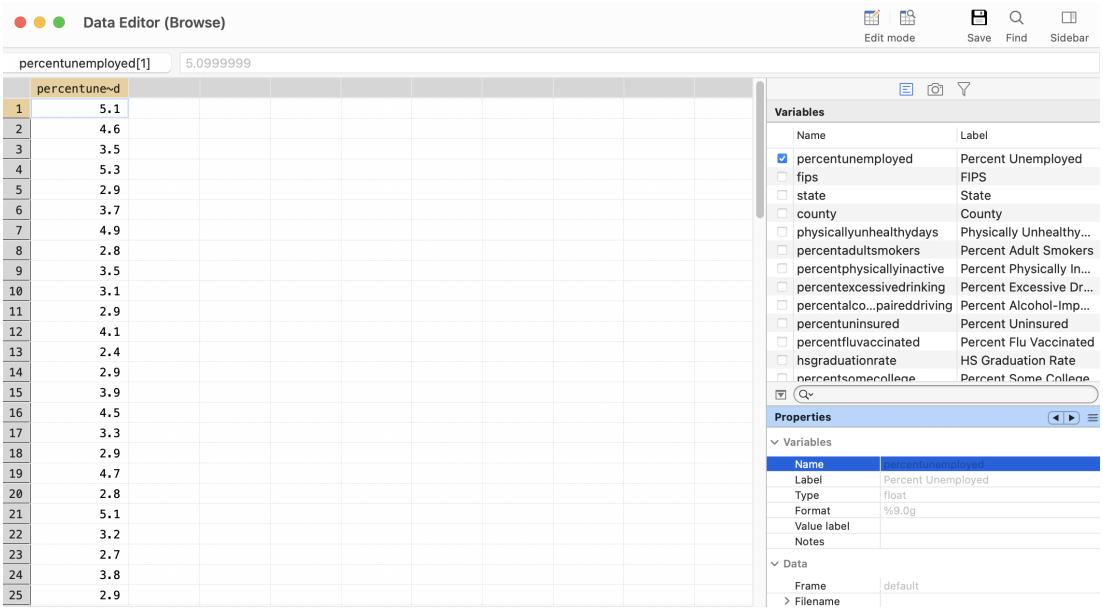
```
Sorted by:
     Note: Dataset has changed since last saved.
```

4. Browse the variable related to Percent Unemployed only.

   From the result of running the `describe` command, the variable name associated with Percent Unemployed is `percentunemployed`. Thus, the command to use for browsing the Percent Unemployed variable is

   ```
   browse percentunemployed
   ```

   Running this line of command pops up a new window that contains only the Percent Unemployed variable for browse. It looks like the following:



5. Create a histogram of Percent Unemployed, and save it as a PNG file called "q5_histogram.png". Is the distribution of Percent Unemployed skewed in any way?

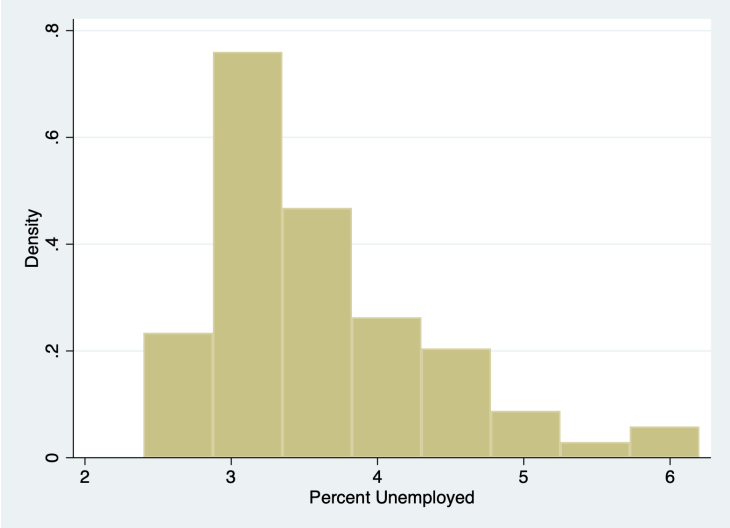
   The command for plotting the histogram is

   ```
   histogram percentunemployed
   ```

   After plotting the histogram, you can either click on the "Save" button in the histogram picture that popped up, and save it on your laptop as "q5_histogram.png". Alternatively, you can type down the following command to save your plot:

   ```
   graph export "q5_histogram.png", replace
   ```

   The plotted histogram looks like the following:



The hisotgram shows that the distribution of Percent Unemployed is positively skewed.

6. What's the mean, variance, and the 10th percentile of Percent Unemployed?

   To obtain such information, we need to summarize the `percentunemployed` variable. Notice that we need information related to percentile, so the `detail` option should be added. Overall, the command to use is

   ```
   summarize percentunemployed, detail
   ```

   The result from running this command is the following:

```
. summarize percentunemployed, detail

                    Percent Unemployed
-------------------------------------------------------------
      Percentiles      Smallest
 1%           2.4           2.4
 5%           2.8           2.5
10%           2.8           2.7       Obs                  72
25%             3           2.8       Sum of wgt.          72

50%           3.4                     Mean                3.6
                        Largest       Std. dev.       .7906139
75%           3.9           5.1
90%           4.7           5.3       Variance        .6250704
95%           5.1             6       Skewness        1.165459
99%           6.2           6.2       Kurtosis        4.228344
```

   From this output, the mean of Percent Unemployed is 3.6, the variance is about 0.6251, and the 10th percentile is 2.8.

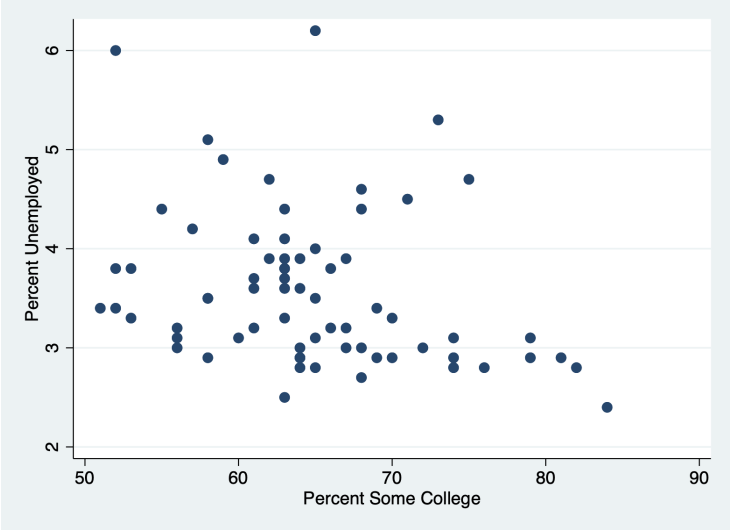
7. Create a scatter plot between Percent Unemployed and Percent Some College for observations where Percent Some College is above 50.

   From question 3, the variable that describes Percent Some College is called `percentsomecollege`. So essentially, we want to create a scatter plot between `percentunemployed` and `percentsomecollege`.

Now, to limit the observations, the `if` syntax can help us with that. Overall, the command to use for this question is

```
scatter percentunemployed percentsomecollege if percentsomecollege > 50
```

The resulting scatter plot from running this command is the following:



8. What is the correlation between Percent Unemployed and Percent Some College for observations where Percent Some College is above 50? Interpret the correlation measure.

   The command to use is

   ```
   correlate percentunemployed percentsomecollege if percentsomecollege > 50
   ```

   The result from running this command is the following:

```
. correlate percentunemployed percentsomecollege if percentsomecollege > 50
(obs=69)

             | perc~yed perce~ge
-------------+------------------
percentune~d |   1.0000
percentsom~e |  -0.2880   1.0000
```

   From the table, the correlation coefficient between Percent Unemployed and Percent Some College for observations where Percent Some College is above 50 is -0.2880. This is a negative number, so the relationship between Percent Unemployed and Percent Some College for the limited observations is negative. The absolute scale of the number is closer to 0 than to 1, so the strength of the relationship is weak.

   Thus, the correlation of coefficient implies that the relationship between Percent Unemployed and Percent Some College for observations where Percent Some College is above 50 is weakly negative.

9. Export your Stata output result as a PDF file.

   At this stage, we've finished all parts of the question. Before exporting the result output, I'd recommend clearing the Results panel following the step in question 1(a) one more time, just to get rid of the potential errors you made while running your code the first time around.

   Now, run the entirety of your Do-file, and export your result by visiting **File** > **Print** > **Results**, and save the results as a PDF using your OS's printing dialogue.

2 Exam 2 Question Revisit

[This part of the handout is released at 5:30pm on Oct 21 (after Exam 2 has closed for all students)]

1. The amount of money spent by UW Madison students at Best Buy in August is a normal random variable with a mean of \$650 and standard deviation of \$150. Find the interquartile range for this variable.

   Recall that interquartile range (IQR) is defined as the following:

   $$IQR = \text{75th percentile} - \text{25th percentile}$$

   Let random variable X be the amount of money spent by UW Madison students at Best Buy in August. Based on the information given in the question,

   $$X \sim N(650, 150^2)$$

   To find the IQR of X, equivalently, we need to find the 75th and 25th percentile of X. Note that, if we denote the 75th percentile of X as X_{75}, then X_{75} must satisfy the following condition:

   $$P(X \leq X_{75}) = 0.75$$

   With X following a normal distribution, we need to standardize X to a standard normal distribution so that we can solve for the value of X_{75}:

   $$P\left(\frac{X - 650}{150} \leq \frac{X_{75} - 650}{150}\right) = 0.75$$
   $$P\left(Z \leq \frac{X_{75} - 650}{150}\right) = 0.75$$

   Based on the z-table, when $P(Z \leq z) = 0.75$, the small z value should be 0.675.

$P(-\infty < Z < z)$

Z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549

(Aside: since we don't have an exact small z value from the z-table that gives us 0.75 probability, technically any small z value between 0.67 and 0.68 would be counted as correct.)

Since we have $P\left(Z \leq \frac{X_{75}-650}{150}\right) = 0.75$ and $P(Z \leq 0.675) = 0.75$. This means that

$$\begin{aligned} \frac{X_{75} - 650}{150} &= 0.675 \\ X_{75} &= 650 + 150 \times 0.675 \\ X_{75} &= 751.25 \\ \Leftrightarrow \quad \text{75th percentile} &= 751.25 \end{aligned}$$

Similarly, we can solve for the 25th percentile. Denote the 25th percentile of X as X_{25}. Then,

$$\begin{aligned} P(X \leq X_{25}) &= 0.25 \\ P\left(\frac{X - 650}{150} \leq \frac{X_{25} - 650}{150}\right) &= 0.25 \\ P\left(Z \leq \frac{X_{25} - 650}{150}\right) &= 0.25 \end{aligned}$$

There are a couple ways that you can find the small z value that corresponds to $P(Z \leq z) = 0.25$. Here, we are going to use the symmetry of the standard normal distribution.

Recall that the standard normal distribution is symmetric around 0. We found earlier that $P(Z \leq 0.675) = 0.75$.

By symmetry, $P(Z > -0.675) = 0.75$, which means that $P(Z \leq -0.675) = 1 - P(Z > -0.675) = 0.25$. Hence,

$$\begin{aligned} \frac{X_{25} - 650}{150} &= -0.675 \\ X_{25} &= 650 - 150 \times 0.675 \\ X_{25} &= 548.75 \\ \Leftrightarrow \quad \text{25th percentile} &= 548.75 \end{aligned}$$

Therefore,

$$\begin{aligned} IQR &= \text{75th percentile} - \text{25th percentile} \\ &= 751.25 - 548.75 = 202.5 \end{aligned}$$

(On the exam, we accept answers that are close to 202.5 to allow for rounding errors.)