Dis 10: Stata Q&A

Check out the solution from Dis 4 (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 Apr 30 @ 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 4) 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**

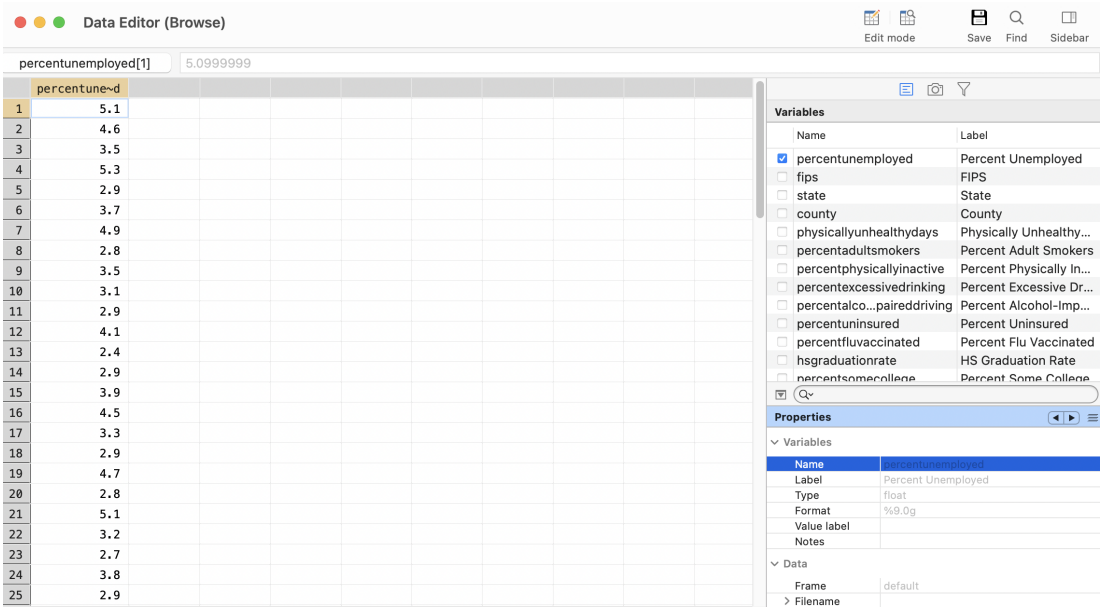
```
Sorted by:
     Note: Dataset has changed since last saved.
```

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?

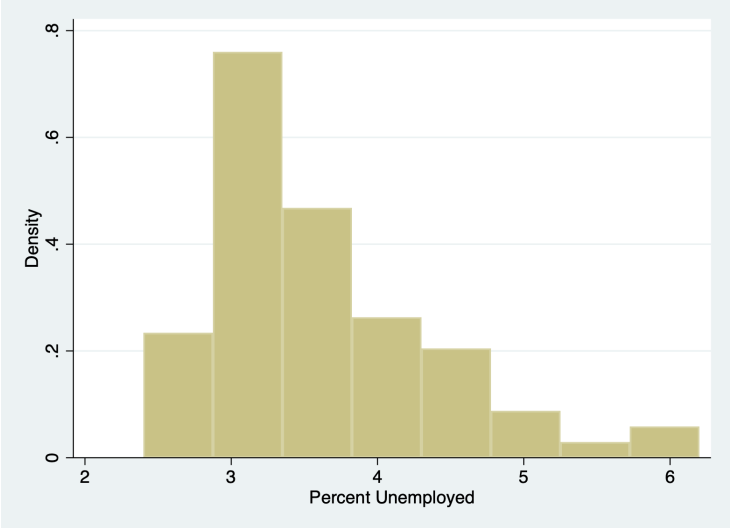
   The command for plotting the histogram is

   ```
   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. Calculate the 93% confidence interval for mean of Percent Unemployed.

   To obtain the 93% confidence interval at the mean, we need the `ci means` function. Here, to specify that the confidence level is 93%, we need to attach the `level(93)` option at the end. Thus, the command to use for this question is

   ```
   ci means percentunemployed, level(93)
   ```

   The result from running this command is the following:

   ```
   . ci means percentunemployed, level(93)
   ```

Variable	Obs	Mean	Std. err.	[93% conf.	interval]
percentune~d	72	3.6	.0931747	3.428591	3.771409

   The last two columns labelled under `[93% conf.interval]` give us the lower bound and the upper bound of the 93% confidence interval. Thus, the 93% confidence interval for mean of Percent Unemployed is

   $$[3.428591, 3.771409]$$

7. 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
```

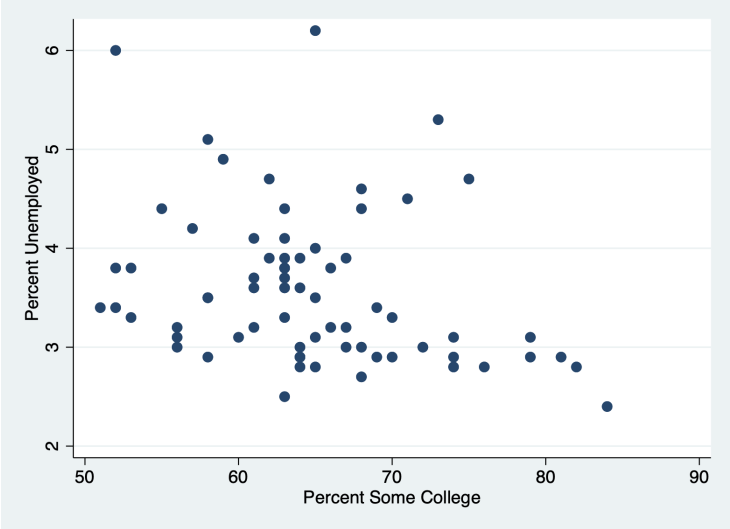
From this output, the mean of Percent Unemployed is 3.6, the variance is about 0.6251, and the 10th percentile is 2.8.

8. 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:



9. 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.

10. 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 A Type-II Error Question from Exam 3

1. The temperature (measured in Fahrenheit) in May in Madison is distributed with standard deviation of 10 F. Suppose that a sample of 31 random days in May over the years has been collected through simple random sampling, and the average May temperature in Madison in your sample is found to be 50.72 F.

   In the first two parts of this question, we tested the following hypothesis:

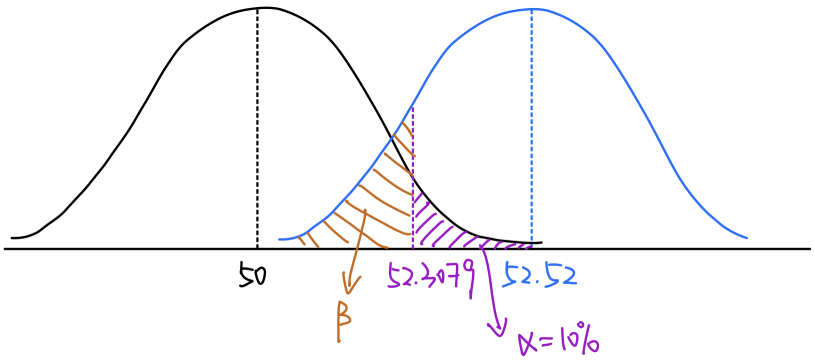
$$H_0 : \mu = 50$$
$$H_1 : \mu > 50$$

   (c) Suppose that the true population average of May temperature in Madison is 52.52 F. What's the probability of committing type-II error for the test you just conducted under 10% significance level?

   Recall when is type-II error committed: type-II error occurs when the null hypothesis is false, but we fail to reject such null hypothesis.

   In this case, our null hypothesis is that $\mu = 50$. Turns out, the true $\mu = 52.52$. Plot both normal distributions for the sample mean $\bar{X}$ (by the way, $\bar{X}$ follows a normal distribution here since

$n = 31 \geq 30$, so central limit theorem gives us an approximated normal distribution), we then need to find the appropriate area where type-II error has been committed.



β labelled on the graph is the probability of committing type-II error, and let's see why. Since our significance level is 10%, we can first find at which level k that we start rejecting the null hypothesis:

$$\begin{aligned} P(\bar{X} > k) &= 0.10 \\ 1 - P(\bar{X} < k) &= 0.10 \\ P(\bar{X} < k) &= 0.90 \\ P\left(Z < \frac{k - \mu_0}{\sigma/\sqrt{n}}\right) &= 0.90 \end{aligned}$$

Looking up the z-table, we find that at $z = 1.285$, $P(Z < z) = 0.90$. This means that

$$\frac{k - \mu_0}{\sigma/\sqrt{n}} = 1.285$$

With $\mu_0 = 50$, $\sigma = 10$, and $n = 31$, we have $k = 52.3079$.

Now, for any sample mean greater than 52.3079, we reject the null hypothesis. Equivalently, for any sample mean less than 52.3079, we fail to reject the null. Given that the true population mean should be 52.52, we can now find the probability that the null wasn't rejected, conditional on $\mu = 52.52$. This will be the probability of committing type-II error (β):

$$\begin{aligned} \beta &= P\left(\bar{X} < 52.3079 | \mu = 52.52\right) \\ &= P\left(Z < \frac{52.3079 - \mu}{\sigma/\sqrt{n}} \middle| \mu = 52.52\right) \\ &= P\left(Z < \frac{52.3079 - 52.52}{10/\sqrt{31}}\right) \\ &= P\left(Z < -0.12\right) = P(Z > 0.12) = 1 - P(Z < 0.12) = 1 - 0.5478 = 0.4522 \end{aligned}$$

Thus, the probability of committing type-II error is 0.4522.