

## Homework 2

---

Homework 1 is due July 13th at 2:00 PM PDT through Canvas. This problem set is worth **50 points**, which we will convert into **20% of your grade**.

You will turn in (i) a write-up with answers to all problems, (ii) a well-commented Stata do file with the solutions for Q2 and Q3, and (iii) a log file generated when executing the do file. All the files should be named as **LastName.FirstName**. There are graph questions in Q3. Include all the graphs in your write-up as graphics.

1. What will be the STATA output of the following commands: **[7 X 2= 14 points]**
  - (i) `dis (21>12)`
  - (ii) `dis 21==12`
  - (iii) `dis 21 !=12`
  - (iv) `dis (43+5 > 12) | (43 > 12-5)`
  - (v) `dis ((43+5 > 12) | (43 > 12-5)) & (12 >21 | 10>12 )`
  - (vi) `dis "2+2"`
  - (vii) `dis "2+2 =" 2+2`
  
2. To answer the following questions, download the dataset named "recent\_grads.xls". This dataset pertains to information about various majors offered in a university. It contains information on enrollment, employment/ unemployment rates and salaries of the graduates of respective majors. The data reports first quartile, third quartile and median of the salary distribution in each major.  
 Change the directory to where you can find your work. Answer the following questions based on this dataset. **[22points]**
  - (i) Import the dataset into stata and save it as a dta file. **[1 points]**
  - (ii) For this question, we only need variables *Major\_code*, *Major*, *Total*, *Men*, *Women*, *Major\_category*, *ShareWomen*, *Unemployment\_rate*, *Median*, *P25th* and *P75th*. Drop the data on all other variables. **[1 points]**
  - (iii) What is the unit of analysis? **[1 points]**
  - (iv) Rename variables- *Total*, *Men*, *Women* and *ShareWomen* respectively as *Enrollment\_Total*, *Enrollment\_Men*, *Enrollment\_Women* and *Share\_Women\_Enroll*. **[4 X 0.5= 2 points]**
  - (v) Label the variable *Share\_Women\_Enroll* as share of women among total students enrolled. **[1 points]**
  - (vi) Create a new variable- *IQ\_range* that captures the Interquartile range (P75th- P25th). Label it as Interquartile Range. **[2 X 1= 2 points]**
  - (vii) What is the major for which *Share\_Women\_Enroll* is missing? **[1 points]**
  - (viii) How many Major categories are there in the dataset? **[1 points]**
  - (ix) Create a new variable- *Science* that takes the value 1 if major categories are *Agriculture & Natural Resources*, *Biology & Life Science*, *Computers & Mathematics*, *Engineering*, *Physical Sciences* and 0 for all other categories. **[2 X 1= 2 points]**
  - (x) What is the mean value of Median salary for Science graduates? What is it for non-Science graduates? **[2 X 1= 2 points]**

- (xi) Calculate the average unemployment rate for Science graduates? What is it for non-Science graduates? [**2 X 1= 2 points**]
  - (xii) Which major offers the highest Median salary? [**1 points**]
  - (xiii) Which major hosts the highest number of women as a fraction of its student body? [**1 points**]
  - (xiv) Is there any major for which share of women enrollees is no less than 70% and the unemployment rate is  $\leq 2\%$ ? [**1 points**]
  - (xv) Debug the following codes, i.e, find and correct the mistake in the following codes: [**3 X 1= 3 points**]
    - i. `tab Major_category if Science= 1`
    - ii. `count if Major_category == Engineering`
    - iii. `count if College_jobs != .`
3. For this question, we will call up one of the in-built datasets in STATA. Please run the following command in STATA:
- sysuse auto.dta, clear*
- This dataset contains information about various features of cars manufactured in 1978. Answer the following questions based on this dataset. [**5 X 2= 10 points**]
- (i) How many observations are there in this dataset? Identify the string variable in the dataset.
  - (ii) Draw a histogram of variable- *price*. Give it the title- “Distribution of price”. Overlay a kdensity function on it. Make sure that the bin size is 500
  - (iii) Now make a box plot of the variable- *price* over car type (*foreign*).
  - (iv) Draw a scatter plot of price (on y-axis) on mileage (on x-axis). Give it the title- “Price v/s Mileage”. Use red dots for *domestic* make and blue dots for *foreign* make.
  - (v) Draw a scatter plot of price (on y-axis) on mileage (on x-axis). Give it the title- “Price v/s Mileage”. Add the best fit line to the plot. Make sure you use blue-colored dots for scatter plot and red color for line.
4. **Making Progress towards the Final Project:** [**2 X 2= 4 points**]
- At this point in the course, we would like to make sure that you are thinking about your final project. The objective of this question is to have you suggest a source of data for your final project so we can give you feedback on the source and the question you might ask.
- If you are having trouble finding a topic, feel free to draw inspiration from the given datasets and choose a dataset from here for your final project.
- (a) Which dataset, from either the provided or your own research, are you interested in working on?
  - (b) What is the question you are interested in studying?