Econ 251

Problem Set #2

(38 points)

Part I: Data analysis (14 points in total)

This part of the problem set introduces you to using STATA for simple data analysis.

Note: Refer to your section 2; you used all of the STATA commands needed in this homework in the section.

Instructions:

Following each question, **please handwrite or type your answers** and **copy/paste the STATA output** (please, use the '**copy as picture**' option). **Your STATA output should be included as part of the homework submission.**

Background: There is evidence in the economic literature that black workers earn less than non-black workers, on average – a phenomenon which has become known as the "racial earnings gap". It's important to understand the reason for this gap, in particular it is important to find out if there is any labour market discrimination against black workers. By the end of this class we are going to see that there is evidence of such discrimination (sometimes referred to as the "unexplained" part of the wage gap) and we'll also see that finding out the extent of discrimination is a very challenging task.

To address this research question we are going to use data on male workers in the U.S. called WAGE2.dta (uploaded on Canvas). This data was used in a publication by Blackburn and Neumark (1992). We are going to start the analyses by looking at **sample averages** for black and non-black workers.

Download the STATA file WAGE2.dta from the CANVAS website. WAGE2.dta contains information on monthly earnings, employment history, education, demographic characteristics, and two test scores for 935 men in year 1980. In particular, it contains the following variables:

wage	monthly earnings (in 1976 USD)
hours	average weekly hours of work
IQ	IQ (intelligence quotient) score
educ	years of education
age	age in years
married	=1 if the person is married
black	=1 if the person is black

1. *(i)* Find the average years of education for everyone in the sample (variable *educ*).

(ii) What are the lowest (minimum) and the highest (maximum) years of education in the sample? *(1 points)*

Hint: Use the STATA command *sum var1*

2. *(i)* How many black men are there in the sample?
   (ii) How many non-black men are there in the sample?
   (iii) What is the percentage of non-black men in the sample?
   (2 points)

Hint: Use the STATA command *tab var1*

3. *(i)* Find the average monthly wage for all men in the sample (variable *wage*).
   (ii) Find the *sample mean* monthly wage for black and non-black workers separately. Do black men in the sample earn more or less than non-black men, on average? *(2 points)*

Hint: Use the STATA command *tab var1, sum(var2)*
In our particular case, *var1* is *black* and *var2* is *wage;* hence, what you would need to type in the command window in STATA is: *tab black, sum (wage)*

(iii) Does the information on the sample mean wages in part *(ii)* above provide compelling evidence that black workers are discriminated against on the labour market (i.e. that they get lower wages *only* because they are black)? Why or why not? *(2 points)*

Hint: This question draws your attention to the difference between *sample statistics* and *population parameters*, and the difference between *correlation* and *causality*.

4. Tabulate variable *married*. Category *married==1* includes people who are married, while category *married==0* includes people who are single, separated, divorced or widowed. How many people fall in category *married==0?* Now find the *sample average* years of education for married and non-married separately. Which group has a higher *sample mean* education? *(2 points)*

Hint: To answer the question use the STATA command *tab var1, sum(var2)* again, but this time *var1* is *married*, *var2* is *educ*.

5. *(i)* What is the *sample correlation* between monthly earnings (*wage*) and years of education (*educ*)? What is the sign of this correlation and what does it mean?

Hint: Use the STATA command *corr var1 var2* *(2 points)*

(ii) Compute the average monthly wage for each year of education in the sample. Do you find any relationship between the two variables? Does this make sense to you? Also, produce a scatter plot of wages and education. Does the scatter plot confirm this relationship? *(3 points)*

Hint: Use the STATA command *tab var1, sum(var2)*
In this case, *var1* is *educ* and *var2* is *wage.*

Part II: Statistical theory (24 points in total)

Note: Please, refer to your lecture notes and notes on section 3 when solving Part II of this homework. These problems are very similar to some examples you saw in class and in your section 2.

Problem 1 (16 points)

You would like to know the average wage of all working women and men between the ages of 18 and 54 in the Ann Arbor area. Suppose the population has mean μ and variance σ^2. You find it too costly in terms of time and money to ask everyone about their earnings, so you *randomly* select 500 people from this group, and ask them about their wage.

(i) What are the *population* and the *population parameter* of interest in this example? What is the *sample*, and what is the *sample size*, N?

(1 point)

Now, comment on each of the proposed *estimators* of the *population mean* wage, μ *:*

(ii) using the sample mean $\bar{X}$ as an *estimator* of the population mean

- Is it *unbiased*? Show formally and explain what this means intuitively.
- Is it *efficient*? Why or why not? Explain intuitively what efficiency means.
- Show formally that its variance equals $\frac{\sigma^2}{N}$.

Hint: we did this in class.

(2 points)

(iii) another *estimator* $\widehat{\mu}_1$ using only three observations from your sample and calculated as: $\widehat{\mu}_1 = \frac{1}{3}(X_1 + X_2 + X_3)$ (the *average* between the first three observations).

- Is it *unbiased*? Show formally.
- Is it *efficient*? Why or why not?
- Show formally that its variance equals $\frac{\sigma^2}{3} > \frac{\sigma^2}{N}$.

(3 points)

(iv) another *estimator* $\widehat{\mu}_2$ using only two observations from your sample and calculated as: $\widehat{\mu}_2 = \frac{1}{3}X_1 + \frac{2}{3}X_2$ (a *weighted average* between the first and second observation with a higher weight placed on the second observation).

- Is it *unbiased*? Show formally.
- Is it *efficient*? Why or why not?
- Show formally that its variance equals $\frac{5}{9}\sigma^2 > \frac{\sigma^2}{N}$.

(3 points)

(v) another *estimator* $\widehat{\boldsymbol{\mu}}_{\boldsymbol{3}}$ using only the first and the last observation from your sample and calculated as:

$$\widehat{\boldsymbol{\mu}}_{\boldsymbol{3}}=\frac{\boldsymbol{X_1+NX_N}}{\boldsymbol{2N+2}}, \text{ where } N \text{ is the sample size.}$$

- Show that this estimator is *biased*, i.e. show that $\mathbb{E}(\widehat{\boldsymbol{\mu}}_{\boldsymbol{3}}) \neq \mu$.
- Find the bias of $\widehat{\boldsymbol{\mu}}_{\boldsymbol{3}}$. Is the estimator biased up or down (i.e. does it overestimate or underestimate μ?
- If it is biased, can it be efficient?

(3 points)

(vi) Rank the estimators in parts *(ii)* to *(v)* in order of your preference starting from the one you would prefer most to the one you would prefer least. Explain.

(3 points)

Hint: Both $\widehat{\mu}_1$ and $\hat{\mu}_2$ are unbiased; in order to choose between these two unbiased estimators, we need to compare their variances (we would prefer the one with a smaller variance). Is $\text{var}(\hat{\mu}_1) > or < \text{var}(\hat{\mu}_2)$?

(vii) Suppose now you take 5,000 samples of size 500 and calculate the sample mean in each sample and plot it on a graph. Approximately what will be the sampling distribution of the sample means? Be as precise as possible. Does your answer depend on the distribution of the population?

(1 points)

Problem 2 (8 points)

Recall the die rolling example from class: you are interested in estimating the population mean of a random variable X describing the outcomes of rolling a 6-sided die (i.e. X follows a uniform distribution). We know that the true population mean of X is 3.5 but suppose we didn't know it and we would like to estimate it.

Consider now the following estimators of the population mean μ:

$$\hat{\mu}_4 = \frac{9}{10}\overline{\text{X}}$$
$$\hat{\mu}_5 = \frac{N-1}{N}\overline{\text{X}}$$

where $\overline{\text{X}}$ is the sample mean and *N* is the sample size.

(i) Show that both $\hat{\mu}_4$ and $\hat{\mu}_5$ are biased and find their biases.

(ii) Now open Excel file HW2 Dice roll.xls. Tab N=100 shows the estimates when using each estimator and a sample of size N=100. Press F9 to generate a new sample and observe the value of the estimates – are the estimates always close to the true population mean 3.5?

Note: If you have a Mac there is no F9 key. To generate a new sample do the following: place the cursor in any cell, type in a number (any number, e.g. 5) and press "Enter". The worksheet will refresh and a new sample will be generated.

(iii) Do the same with Tab N=50,000, which shows the estimates when using each estimator and a sample of size N=50,000. What happens to $\hat{\mu}_4$ as the sample size gets large? What about $\hat{\mu}_5$? Which one of the estimators gets closer to the true population mean 3.5?

(iv) What important property of estimators does this exercise illustrate?

(2 points each, 8 points in total)