Sample 1: Midterm Solutions

ECON 340: Economic Research Methods — Instructor: Div Bhagia

Print Name: _____

This is a closed-book test. You may not use a phone or a computer.

Time allotted: 70 minutes
Total points: 20

Please show sufficient work so that the instructor can follow your work.

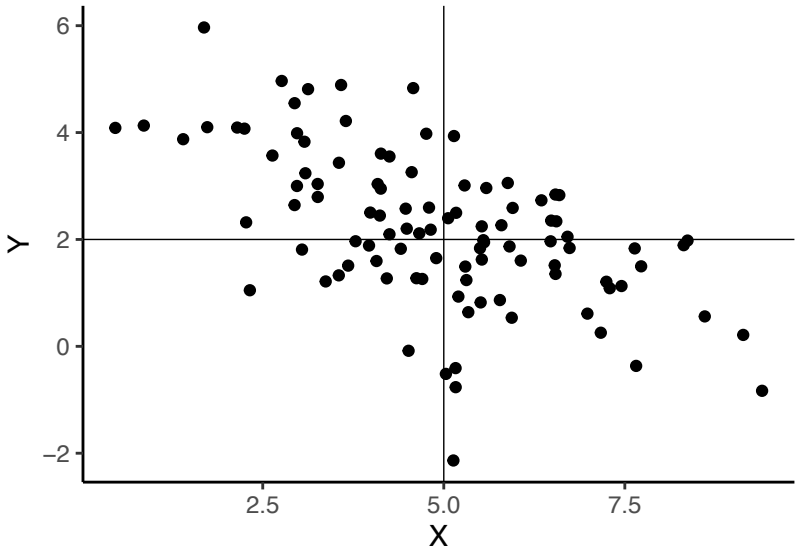
I understand and will uphold the ideals of academic honesty as stated in the honor code.

Signature: _____

Question 1: Multiple Choice Questions (1 pt each, total 5 pts)

Select one correct answer unless specified otherwise.

(a). Below is a scatter plot of two variables X and Y.



What is the correlation between these variables?

- ☐ $\rho = 0.4$
- ☐ $\rho = -1$
- ☑ $\rho = -0.6$
- ☐ $\rho = 0$

(b). If the average height in the world is 169 cm with a standard deviation of 6 cm, and my height is 167 cm, how many standard deviations below the mean am I?

- ☐ 1.24
- ☑ 0.33
- ☐ 0.67
- ☐ 2.82

(c). X and Y are two *independent* random variables. X is a binary variable that takes the value 1 or 0, and Y is a continuous variable. We know that $E(Y|X = 1) = 100$, then:

- ☐ $E(Y|X = 0) = 50$
- ☑ $E(Y|X = 0) = 100$
- ☐ $E(Y|X = 0) = 200$
- ☐ We can't say anything about $E(Y|X = 0)$ with the above information.

(d). Sample mean $\bar{X}$ is an unbiased estimator for the true population mean μ. Does this mean

- ☐ $\bar{X} = \mu$ in all samples.
- ☑ $\bar{X} = \mu$ on average across samples.
- ☐ $E(\bar{X}) = 0$
- ☐ All of the above

(e). (Select all that apply.) Which of the following statements about confidence intervals is true?

- ☐ The confidence interval widens as the population variance decreases.
- ☑ The confidence interval widens as the population variance increases.
- ☑ The confidence interval widens as the sample size decreases.
- ☐ The confidence interval widens as the sample size increases.

Question 2: Things you can explain (6 pts)

(a). (2 pts) *In most countries across the world mean earnings are greater than median earnings. Why do you think this is true?*

In most countries around the world, average earnings exceed median earnings. This discrepancy arises because a small fraction of people earn exceptionally high salaries, skewing the average upwards. For instance, in the United States, the earnings of the top 5% account for 28% of total earnings. [1] If we plot a histogram of earnings, we will notice that the distribution of earnings is right-skewed. The mean is affected by such outliers as it depends on all values in the data, while the median is less sensitive to extreme values as it only depends on values in the center of the data.

(b). (2 pts) *There is a strong positive correlation between immigrant flows and online job vacancies across metropolitan areas in the US. Does this imply that immigration leads to job creation?*

No, this does not imply that immigration leads to job creation because correlation does not imply causation. Immigrants may choose to move to places in response to job vacancies, resulting in a positive correlation between immigrant flows and online job vacancies. This is called reverse casualty.

Alternative answer: No, this does not imply that immigration leads to job creation because correlation does not imply causation. Immigrants may settle in areas with [*insert characteristic*] and areas with [*insert characteristic*] also attract more jobs which would result in a positive correlation between immigrant flows and online job vacancies.

[1] Estimates from Social Security Data for the year 2017.

(c). (2 pts) *As long as our sample is random (no matter how small or large), the sample mean is going to give us an unbiased estimate of the true population mean. Why do we still care about having a large sample?*

We care about having a large sample because the variance of the sample mean given by σ^2/n is lower when the sample size n is greater. Unbiasedness implies that, on average, we will get a sample mean equal to the true population mean. It is still very possible to get a sample mean far from the true mean in any given sample; how far depends on the variance of our estimate. Another benefit of having a large sample is that by the Central Limit Theorem, the sample mean distribution in large samples is normal even when the population is non-normal. Knowing the sample mean distribution enables us to construct confidence intervals and perform hypothesis tests.

Question 3: Voter Turnout (9 pts)

We took a random sample of 200 individuals from the US population and asked them whether they voted in the last election. We then create a variable X_i that takes value 1 if the individual voted in the last election and 0 if they did not. 140 individuals said they voted in the last election, while 60 individuals said they didn't. This data is presented in the table below.

(1) X_i	(2) n_i	(3) f_i	(4) $f_i X_i$	(5) $(X_i - \bar{X})^2$	(6) $f_i(X_i - \bar{X})^2$
1	140	**0.7**	**0.7**	**0.09**	**0.063**
0	60	**0.3**	**0**	**0.49**	**0.147**
Total	200	**1**	**0.7**	NA	**0.21**

(a). (2 pts) *Fill in the third and the fourth column of the above frequency distribution table and calculate $\bar{X}$, the sample mean of X_i. Specify the formula you will use and then plug in the values to calculate your answer.*

$$\bar{X} = \sum_{i=1}^{k} f_i X_i = 0.7$$

(b). (2 pts) *Your answer in (a) tells us the proportion of individuals in our sample who voted in the last election. Can we directly infer the percentage of the US population that voted in the last election from this answer? Explain.*

We cannot directly say that 70% of the US population voted in the last election. The sample mean is a random variable, and depending on the sample we picked, we could be further or closer to the actual population average. However, since we took a large random sample, the sample mean is normally distributed around the true population mean. So we can construct confidence intervals and perform hypothesis tests to infer something about the true population mean.

(c). (2 pts) *Fill in the fifth and the sixth column of the above frequency distribution table and calculate S_X^2, the sample variance of X_i. Specify the formula you will use and then plug in the values to calculate your answer.*

$$S_X^2 = \frac{n}{n-1}\sum_{i=1}^{k} f_i(X_i - \bar{X})^2 = \frac{200}{199} \cdot 0.21 = 0.211$$

(d). (2 pts) Construct a 90% confidence interval for voter turnout in the U.S. population.

Note: $Pr(Z > 1.645) = 0.05$, $Pr(Z > 1.96) = 0.025$, $Pr(Z > 2.33) = 0.01$

Here, $1 - \alpha = 0.9$, so the critical value we need to use for the confidence interval is $t_{199,0.05}$. Since the t distribution with large degrees of freedom can be approximated using the standard normal distribution, we have $t_{199,0.05} \approx z_{0.05} = 1.645$. Therefore, our confidence interval is calculated as follows:

$$\bar{x} \pm t_{199,0.05} \cdot \frac{S}{\sqrt{n}} = 0.7 \pm 1.645 \cdot \sqrt{\frac{0.211}{200}}$$

So the 90% confidence interval for voter turnout is given by: [0.647, 0.753].

(e). (1 pt) Interpret the confidence interval you've constructed in part (d). What does it reflect in this case?

The confidence interval implies that we are 90% confident that between 64.7% to 75.3% of the population voted in the last election.