Spring 2023 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)

Choose a single correct response for all questions except for (a), which allows for the selection of multiple answers.

(a). (Select all that apply) Suppose a survey of 500 people found that 300 of them prefer coffee over tea. Suppose we define a variable X that takes a value of 1 if a person said they prefer coffee and 0 otherwise. The sample mean of X is given as:

- ☑ $1/500(1 \cdot 300 + 0 \cdot 200)$
- ☑ $(1/500)\sum_{i=1}^{500} X_i$
- ☑ $1 \cdot (3/5) + 0 \cdot (2/5)$
- □ None of the above

(b). Which of the following statements explains why a random sample is preferred when calculating the sample mean?

- □ A random sample ensures that the sample mean equals the population mean.
- □ A random sample reduces the variance of the sample mean.
- ☑ A random sample avoids bias in the selection of individuals for the sample.
- □ A random sample guarantees that the sample mean is normally distributed.

(c). Which of the following statements correctly explains the concept of independence between two random variables?

- □ Two random variables are independent if they have no outcomes in common.
- □ Two random variables are independent if they have exactly the same outcomes.
- ☑ Two random variables are independent if the value of one random variable does not affect the probability distribution of the other random variable.
- □ Two random variables are independent if they have the same probability distribution.

(d). If the average score on a math test is 75 with a standard deviation of 5, and your score is 85, how many standard deviations above the mean are you?

- ☐ 1
- ☑ 2
- ☐ 3
- ☐ 4

(e). Which of the following statements correctly explains the concept of p-values in hypothesis testing?

- ☐ The p-value is the probability that the null hypothesis is true.
- ☐ The p-value is the probability of observing the null hypothesis given the sample data.
- ☑ The p-value is the probability of observing the sample data or more extreme values, assuming the null hypothesis is true.
- ☐ The p-value is the probability of observing a statistically significant result in a hypothesis test.

Question 2: Things you can explain. (5 pts)

(a). (2.5 pts) Suppose that in a certain population, 50% of likely voters are women. If a survey is conducted using a random sample of 1000 landline telephone numbers and the results show that 57% of respondents are women, is there any indication that the survey may be biased? Explain.

The answer I was looking for:

Based on the information provided, we cannot conclude that the survey is biased. This is because while we expect the sample mean from a random sample to be equal to the population mean on average, there can be variation across different samples. Therefore, it is possible to obtain a sample mean of 0.57, even if the population mean is 0.5, without any bias in the survey.

An alternative (almost correct) answer that also got full credit:

While the survey may have been random in selecting the phone numbers, there may be some bias in who owns or picks up the phone. If women are more likely to own or answer the phone, the survey may be biased in favor of women, even if the sampling process was random.

(b). (2.5 pts) If a study finds a strong positive correlation between the number of houses and house prices across US cities, can we conclude that more housing supply leads to higher house prices? Why or why not?

If we observe a strong positive correlation between the number of houses and house prices across US cities, it does not necessarily imply that an increase in housing supply causes higher house prices. In fact, it is possible that higher house prices lead to more housing supply, as builders try to maximize their profits by building in more profitable locations, resulting in a positive correlation between the stock of housing and prices. This is known as reverse causality.

Alternative correct answer:

It is possible that external confounding factors are responsible for the observed positive correlation between housing stock and housing prices. For instance, if certain cities have more desirable amenities, more people may want to live there, leading to higher demand for housing. This, in turn, can result in higher prices and increased housing stock as builders construct more houses to meet the growing demand.

Question 3: Exercise and Sleep. (10 pts)

Let X be the average hours of sleep per day you got last week, and let Y be the average hours you exercised per day last week. You want to look at the relationship between these two variables over the last three weeks.

Note: *Use the population formulas. You can use the table and fill in values or directly apply the formulas. In both cases, write down the formula you are using and show your work.*

Week	X_i	Y_i	$(X_i - \mu_X)$	$(Y_i - \mu_Y)$	$(X_i - \mu_X)^2$	$(Y_i - \mu_Y)^2$	$(X_i - \mu_X)(Y_i - \mu_Y)$
1	6	0.5	-2	-0.1	4	0.01	0.2
2	9	0.3	1	-0.3	1	0.09	-0.3
3	9	1	1	0.4	1	0.16	0.4
	24	**1.8**	**0**	**0**	**6**	**0.26**	**0.3**

(a). (2 pts) Calculate the variance of X and Y.

First, let's calculate the mean of X and Y.

$$\mu_X = \frac{1}{3}\sum_{i=1}^{3} X_i = \frac{24}{3} = 8 \qquad \mu_Y = \frac{1}{3}\sum_{i=1}^{3} Y_i = \frac{1.8}{3} = 0.6$$

Now, we can calculate the variance of X and Y as follows:

$$\sigma_X^2 = \frac{1}{3}\sum_{i=1}^{3}(X_i - \mu_X)^2 = \frac{6}{3} = 2 \qquad \sigma_Y^2 = \frac{1}{3}\sum_{i=1}^{3}(X_i - \mu_X)^2 = \frac{0.26}{3} = 0.0867$$

(b). (1 pt) In the formula for variance, why do we use squared deviations from the mean instead of just using deviations from the mean?

The sum of deviations from the mean is zero, as negative deviations cancel out positive deviations. However, by squaring the deviations from the mean, we give equal weight to both positive and negative deviations, which provides a more accurate measure of the spread or variability of the data.

Also correct: *Squaring (say, instead of just taking absolute deviations from the mean) puts greater emphasis on larger deviations. This means that deviations further from the mean contribute more to the overall variance than those closer to the mean.*

(c). (2 pts) Calculate the covariance between X and Y, denoted by σ_{XY}.

$$\sigma_{XY} = \frac{1}{3}\sum_{i=1}^{3}(X_i - \mu_X)(Y_i - \mu_Y) = \frac{0.3}{3} = 0.1$$

(d). (2 pts) In class, we learned that covariance is positive when two variables move together, meaning that they increase or decrease together. Can you explain how the formula you used in (c) ensures that this is the case?

If we look at the formula for covariance, we can see that it involves summing up the product of deviations from the mean for two variables, X and Y. Whenever both X and Y are above their respective means, the product of their deviations will be positive, and it will contribute a positive number to the overall sum. Likewise, whenever both are below their respective means, the product will be positive and will contribute a positive number to the sum. In contrast, when one variable is above its mean while the other is below its mean, the product of deviations will be negative, contributing a negative number to the sum.

Thus, the sign of the covariance indicates whether the two variables move together or in opposite directions. A positive covariance means that on average, when one variable is above its mean, the other variable tends to be above its mean as well, and when one is below its mean, the other tends to be below its mean as well. In contrast, a negative covariance means that on average, when one variable is above its mean, the other variable tends to be below its mean, and vice versa.

Now say instead of recording the exercise in hours, you had recorded it in minutes. Then your data would look as below, where Z is the average minutes of exercise per day.

Week	X_i	Z_i	$(X_i - \mu_X)$	$(Z_i - \mu_Z)$	$(X_i - \mu_X)^2$	$(Z_i - \mu_Z)^2$	$(X_i - \mu_X)(Z_i - \mu_Z)$
1	6	30	-2	-6	4	36	12
2	9	18	1	-18	1	324	-18
3	9	60	1	24	1	576	24
	24	**108**	**0**	**0**	**6**	**936**	**18**

(e). (1 pt) Calculate the covariance between X and Z, denoted by σ_{XZ}.

The mean of Z is given by:

$$\mu_Z = \frac{1}{3}\sum_{i=1}^{3} Z_i = \frac{108}{3} = 36$$

Note that 36 minutes is 36/60=0.6 hours which was the mean of X. So you could have alternatively found the mean of Z as $\mu_Z = 60\mu_X$.

We can calculate the covariance as follows:

$$\sigma_{XZ} = \frac{1}{3}\sum_{i=1}^{3}(X_i - \mu_X)(Z_i - \mu_Z) = \frac{18}{3} = 6$$

(f). (1 pt) Why do you think it is the case that $\sigma_{XZ} > \sigma_{XY}$?

This is because the scale in which a variable is measured can affect the magnitude of the covariance. When we express exercise time in minutes instead of hours, the deviations from the mean are larger in magnitude because each hour is equivalent to 60 minutes, which results in a larger covariance. However, it's important to note that this does not change the sign of covariance, which indicates the direction of the relationship between hours of sleep and exercise time.

(g). (1 pt) There is an alternative statistic that you could calculate, which would ensure that you accurately capture that the relationship between X and Y is as strong as

the relationship between X and Z. What is this statistic?

The correlation is an alternative statistic that can be calculated to accurately capture the strength of the relationship between X and Y relative to the relationship between X and Z. It measures both the direction and the magnitude of the relationship between two variables by standardizing the covariance by the respective standard deviations. So in this case, it will take into account the larger variance of Z relative to X.

Extra credit (1 pt): Calculate this statistic for X and Y and for X and Z.

First note that the variance of Z is given by:

$$\sigma_Z^2 = \frac{1}{3}\sum_{i=1}^{3}(Z_i - \mu_Z)^2 = \frac{936}{3} = 312$$

We can calculate the two correlations as follows:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{0.1}{\sqrt{2} \cdot \sqrt{0.087}} = 0.24$$

$$\rho_{XZ} = \frac{\sigma_{XZ}}{\sigma_X \sigma_Z} = \frac{6}{\sqrt{2} \cdot \sqrt{312}} = 0.24$$