Sample 2: 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) and (b), which allow for the selection of multiple answers.

(a) (Select all that apply) In a survey of 200 people about which film they believed would win the Oscar for Best Picture, 150 respondents predicted Oppenheimer. Let's define a variable X that takes a value of 1 if a person said they thought Oppenheimer will win and 0 otherwise. The sample mean of X is given as:

☑ $(1/200) \sum_{i=1}^{200} X_i$

☑ $1/200(1 \cdot 150 + 0 \cdot 50)$

☐ $1 \cdot (1/5) + 0 \cdot (4/5)$

☐ None of the above

(b) (Select all that apply) In an economic analysis, researchers are exploring the relationship between X, which represents the number of years of education, and Y, the annual income in US dollars. Additionally, they create a new variable Z to measure annual income in euros. Which of the following is true?

☐ $\text{Corr}(X, Y) \neq \text{Corr}(X, Z)$

☑ $\text{Corr}(X, Y) = \text{Corr}(X, Z)$

☑ $\text{Cov}(X, Y) \neq \text{Cov}(X, Z)$

☐ $\text{Cov}(X, Y) = \text{Cov}(X, Z)$

Note: Corr represents correlation, and Cov stands for covariance.

(c) In the class, we learned that the sample mean $\bar{X}$ is an unbiased estimator of the population mean μ, meaning $E(\bar{X}) = \mu$. This means that

☐ In very large samples, the sample mean will converge to the population mean.

☑ The average of sample means, obtained from repeated samples, equals the population mean.

☐ The sample mean is equal to the population mean in most samples.

☐ The variance of the sample mean is very small.

(d) What does the variance of the sample mean signify?

☐ Dispersion in sample data

☑ Sample-to-sample variability

☐ Spread of the population data

☐ Bias in sample mean

(e) Which of the following is an accurate interpretation of the result of a hypothesis test with a p-value of 0.03?

☐ There is a 3% probability that the null hypothesis is true.

☐ We have 97% confidence that the alternative hypothesis is true.

☑ If the null hypothesis was true, there is a 3% probability of observing data at least as extreme as what was observed.

☐ If the null hypothesis was not true, there is a 3% probability of observing data at least as extreme as what was observed.

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

(a) (2.5 pts) If a study finds a strong correlation between the number of years spent in formal education and individuals' wage income, can we reasonably conclude that education leads to higher earnings? Explain your reasoning.

No, we cannot conclude that education directly leads to higher earnings, as correlation does not imply causation. Various factors may influence both the level of education an individual attains and their income. For instance, individuals pursuing higher education may differ significantly from those who don't in ambition, discipline, or access to networks, which could contribute to higher earnings.

(b) (2.5 pts) In a survey about household income, only 80% of households respond, with the majority of non-response coming from individuals without college degrees. How does this bias our estimate of household income? How can we address this bias to ensure that our estimate represents the entire population?

Given individuals with higher education usually earn more, having too many of them in our survey could make us overestimate household income. To fix this, we can reweight our sample based on education level, so it mirrors the actual education spread in the population. Specifically, we can assign a higher weight to households without college degrees, making up for their underrepresentation in the survey data.

Question 3: Investor Behavior (6 pts)

Imagine you're a financial advisor studying two types of investors: *risk-averse* (denoted by $X = 1$) and *risk-taking* (denoted by $X = 2$). You're interested in the amount of money Y they are willing to invest in a new startup: either $10,000, $20,000, or $30,000.

The joint probability distribution table below shows the likelihood of an investor from each risk category investing a particular amount in a new startup:

	Investment Amount (Y)			Total
	$10,000	$20,000	$30,000	
Risk-averse ($X = 1$)	0.35	0.15	0.10	0.60
Risk-taking ($X = 2$)	0.05	0.15	0.20	0.40
Total	0.40	0.30	0.30	1

Answer the following questions. *For each question write down the formula you are using if any and show your work clearly.*

(a) (1 pt) What proportion of all investors invest $10,000?

$$Pr(Y = 10000) = 0.4$$

(b) (1 pt) What is the expected amount of investment across all investors?

$$E(Y) = \sum_y y Pr(Y = y) = 0.4(10000) + 0.3(20000) + 0.3(30000)$$

$$= 4000 + 6000 + 9000 = \$19,000$$

(c) (1 pt) What proportion of investors are risk-averse?

$$Pr(X = 1) = 0.6$$

(d) (1 pt) What *proportion* of risk-averse investors invest $10,000?

$$Pr(Y = 10000|X = 1) = \frac{Pr(Y = 10000, X = 1)}{Pr(X = 1)} = \frac{0.35}{0.60} = 0.5833$$

(e) (1 pt) What is the expected amount of investment from a risk-averse investor?

$$E(Y|X = 1) = \sum_y yPr(Y = y|X = 1)$$
$$= \frac{0.35}{0.60}(10000) + \frac{0.15}{0.60}(20000) + \frac{0.10}{0.60}(30000)$$
$$= 5833.33 + 5000 + 5000 = \$15,833.33$$

(f) (1 pt) Given your answers above, do you think the investment amount is independent of an individual's risk appetite?

No individual's risk appetite is not independent of the investment amount as $E(Y|X) \neq E(Y)$.

Question 4: Estimating Average Income (4 pts)

A university wants to test whether online classes yield similar exam scores to in-person classes. Historically, the average exam score for traditional classes has been 75.4. A random sample of 100 exam scores from online classes reveals an average score of 71.48 with a standard deviation of 20.

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

(a) (2 pts) Conduct a hypothesis test to determine if there is a significant difference in scores between online classes and in-person classes at a 10% level of significance. Formulate the null and alternative hypotheses, clearly define the rejection region, calculate the test statistic, and specify whether you reject or fail to reject the null hypothesis.

We are given, $\bar{x} = 71.48$, $n = 100$, $S = 20$, and $z_{0.05} = 1.645$.
$H_0 : \mu = 75.4$
$H_1 : \mu \neq 75.4$

Test statistic under the null:

$$T = \frac{\bar{x} - \mu}{S/\sqrt{n}} \sim t_{99} \approx N(0, 1)$$

So we will reject the null hypothesis if $|t| > 1.645$.
Calculating the test statistic:

$$t = \frac{71.48 - 75.4}{20/\sqrt{100}} = -1.96$$

Since $|-1.96| > 1.645$, we reject the null hypothesis.

(b) (1 pt) What is the p-value associated with your hypothesis test?

$$p = 2Pr(Z > |t|) = 2Pr(Z > 1.96) = 2 \times 0.025 = 0.05$$

(c) (1 pt) Provide a brief explanation of the logic behind hypothesis testing?

Hypothesis testing uses sample data to evaluate the likelihood that a hypothesis about the population is true. It calculates the probability of observing the sample data if the hypothesis were true and then making a decision on whether to accept or reject the hypothesis based on predefined significance levels.