EC 282 — Midterm 1 Mock Exam

ANSWER KEY

Introduction to Econometrics

Spring 2026

Professor Altindag

Part I: Multiple Choice (3 points each, 45 points total)

1. A **random variable** is best described as:

   (b) A numerical summary of a random outcome

   A random variable assigns a numerical value to each outcome of a random process. It is not always unknown (we observe realized values), does not need to be normal, and is a population concept—not a sample statistic.

2. Let Y be a Bernoulli random variable with $\Pr(Y = 1) = 0.3$. What is $\text{Var}(Y)$?

   (c) 0.21

   For a Bernoulli: $\text{Var}(Y) = p(1 - p) = 0.3 \times 0.7 = 0.21$.

3. Two random variables X and Y are **independent** if and only if:

   (b) $\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$ **for all** x, y

   This is the definition of independence. Note that $\text{Cov}(X, Y) = 0$ is *necessary* but not *sufficient* for independence (zero covariance does not imply independence).

4. If $\text{Cov}(X, Y) = 0$, which of the following is true?

   (b) X and Y have no linear relationship, but may have a nonlinear one

   Zero covariance means no *linear* relationship. However, X and Y could still have a strong nonlinear relationship (e.g., $Y = X^2$ when X is symmetric around zero).

5. Which of the following is **NOT** a property of correlation?

   (c) $\rho_{XY} = 0$ implies X and Y are independent

   Zero correlation means no *linear* relationship, but does not imply independence. All other options are true properties of correlation.

6. If $Y_1, Y_2, \ldots, Y_n$ are i.i.d. with $E[Y_i] = \mu_Y$ and $\text{Var}(Y_i) = \sigma_Y^2$, then $\text{Var}(\bar{Y})$ equals:

   (c) σ_Y^2/n

   $\text{Var}(\bar{Y}) = \text{Var}\big(\frac{1}{n}\sum Y_i\big) = \frac{1}{n^2} \cdot n\sigma_Y^2 = \frac{\sigma_Y^2}{n}$. Note: $\sigma_Y/\sqrt{n}$ is the *standard deviation* (standard error), not the variance.

7. The **Central Limit Theorem** states that:

   (c) For large n, $\bar{Y}$ is approximately normally distributed regardless of the population distribution

   The CLT is about the sampling distribution of $\bar{Y}$, not about the population. It holds regardless of the shape of the original population distribution, which is what makes it remarkable.

8. The **Law of Large Numbers** tells us that:

   (b) $\bar{Y} \xrightarrow{p} \mu_Y$ as $n \to \infty$

   The LLN says the sample mean converges in probability to the population mean as sample size grows. It does not say they are *equal* for any finite n, and it is distinct from the CLT (which is about normality).

9. An estimator $\hat{\mu}_Y$ is **unbiased** if:

   (b) $E[\hat{\mu}_Y] = \mu_Y$

   Unbiasedness means the expected value of the estimator equals the true parameter. It does not mean every individual estimate hits the target—only that on average it does. Option (d) describes *consistency*.

10. The sample variance formula uses $n - 1$ in the denominator because:

    (b) It produces an unbiased estimator of the population variance (Bessel's correction)

    Dividing by $n-1$ corrects for the fact that the sample mean is used in place of the population mean, yielding an unbiased estimator: $E[S_Y^2] = \sigma_Y^2$.

11. The **conditional expected value** $E[Y \mid X = x]$ is:

    (b) The expected value of Y calculated using the conditional distribution of Y given $X = x$

    By definition, $E[Y \mid X = x] = \sum_i y_i \cdot \Pr(Y = y_i \mid X = x)$. It uses the conditional (not marginal) distribution and is generally *not* equal to $E[Y]$ unless X and Y are independent.

12. Suppose X and Y are independent. Which of the following must be true?

    (b) Cov$(X, Y) = 0$

    If X and Y are independent, then $\text{Cov}(X, Y) = 0$. Independence says nothing about whether $E[X] = E[Y]$ or whether the variances are equal. Note: the converse is not true—zero covariance does not imply independence.

13. A population has mean $\mu_Y = 200$ and variance $\sigma_Y^2 = 400$. If you draw a random sample of $n = 100$, then by the CLT the sampling distribution of $\bar{Y}$ is approximately:

    (b) $N(200, 4)$

    By the CLT: $\bar{Y} \overset{a}{\sim} N(\mu_Y, \sigma_Y^2/n) = N(200, 400/100) = N(200, 4)$. The variance of $\bar{Y}$ is $\sigma_Y^2/n = 4$, not $\sigma_Y^2 = 400$.

14. Among two unbiased estimators of μ_Y, the one with **smaller variance** is said to be:

    (b) More efficient

    Efficiency compares the variances of unbiased estimators. The one with smaller variance is more efficient—it produces estimates that are more tightly clustered around the true parameter.

15. A researcher surveys students only from their own lecture section to estimate the average GPA of all students at the university. This is an example of:

    (c) Selection bias (non-random sampling)

    The sample is not randomly drawn from the population of interest. Students in one lecture section may differ systematically from the broader student body, leading to biased estimates.

Part II: Short Answer Problems

Problem 1. Joint Distribution, Conditional Probability, and Covariance (25 points)

(a) **Marginal distributions:**

	$Y = 0$	$Y = 1$	Marginal of X
$X = 0$ (Off-campus)	0.20	0.25	**0.45**
$X = 1$ (On-campus)	0.15	0.40	**0.55**
Marginal of Y	**0.35**	**0.65**	1.00

Marginal of X: $\Pr(X = 0) = 0.20 + 0.25 = 0.45$, $\Pr(X = 1) = 0.15 + 0.40 = 0.55$.

Marginal of Y: $\Pr(Y = 0) = 0.20 + 0.15 = 0.35$, $\Pr(Y = 1) = 0.25 + 0.40 = 0.65$.

(b) **Expected values and variances:**

Since X and Y are Bernoulli:

$$\mathbf{E[X]} = 0(0.45) + 1(0.55) = \mathbf{0.55} \qquad \mathbf{E[Y]} = 0(0.35) + 1(0.65) = \mathbf{0.65}$$

$$\mathbf{Var(X)} = E[X](1 - E[X]) = 0.55(0.45) = \mathbf{0.2475}$$

$$\mathbf{Var(Y)} = E[Y](1 - E[Y]) = 0.65(0.35) = \mathbf{0.2275}$$

Interpretation: 55% of students live on campus, and 65% have a GPA of 3.0 or above.

(c) **Conditional probabilities:**

$$\Pr(Y = 1 \mid X = 1)\Pr(Y = 1 \mid X = 1)\Pr(Y = 1 \mid X = 1)\Pr(Y = 1 \mid X = 1) = \frac{\Pr(X=1, Y=1)}{\Pr(X=1)} = \frac{0.40}{0.55} \approx \mathbf{0.73}$$

$$\Pr(Y = 1 \mid X = 0)\Pr(Y = 1 \mid X = 0)\Pr(Y = 1 \mid X = 0)\Pr(Y = 1 \mid X = 0) = \frac{\Pr(X=0, Y=1)}{\Pr(X=0)} = \frac{0.25}{0.45} \approx \mathbf{0.56}$$

On-campus students have a higher probability (73%) of a high GPA compared to off-campus students (56%).

(d) **Independence test:**

X and Y are independent if $\Pr(X = x, Y = y) = \Pr(X = x) \cdot \Pr(Y = y)$ for *all* x, y.

Check one cell: $\Pr(X = 1, Y = 1) = 0.40$, but $\Pr(X = 1) \cdot \Pr(Y = 1) = 0.55 \times 0.65 = 0.3575$.

Since $0.40 \neq 0.3575$, **X and Y are NOT independent.**

Alternatively: $\Pr(Y = 1 \mid X = 1) = 0.73 \neq 0.65 = \Pr(Y = 1)$, so knowing X changes the probability of Y.

(e) **Covariance and correlation:**

$$\begin{aligned}\text{Cov}(X,Y) &= E[(X-\mu_X)(Y-\mu_Y)] \\ &= \sum_x \sum_y (x-0.55)(y-0.65)\cdot \Pr(X=x, Y=y)\end{aligned}$$

Computing each term:

$$\begin{aligned}&(0-0.55)(0-0.65)(0.20) = (-0.55)(-0.65)(0.20) = 0.0715 \\ &(0-0.55)(1-0.65)(0.25) = (-0.55)(0.35)(0.25) = -0.0481 \\ &(1-0.55)(0-0.65)(0.15) = (0.45)(-0.65)(0.15) = -0.0439 \\ &(1-0.55)(1-0.65)(0.40) = (0.45)(0.35)(0.40) = 0.0630\end{aligned}$$

$$\textbf{Cov(X,Y)} = 0.0715 - 0.0481 - 0.0439 + 0.0630 = \mathbf{0.0425}$$

$$\sigma_X = \sqrt{0.2475} = 0.4975, \qquad \sigma_Y = \sqrt{0.2275} = 0.4770$$

$$\textbf{Corr(X,Y)} = \frac{0.0425}{0.4975 \times 0.4770} = \frac{0.0425}{0.2373} \approx \mathbf{0.18}$$

The positive correlation indicates that on-campus housing and higher GPA tend to go together. However, the magnitude is moderate, and **correlation does not imply causation**—other factors (e.g., income, year in school) could explain this relationship.

Problem 2. Sampling Distribution and Estimation (15 points)

Given: $\mu_Y = 950$, $\sigma_Y = 300$, $n = 225$.

(a) **Expected value of $\bar{Y}$:**

$$\mathbf{E}[] = \mu_Y = 950\mathbf{E}[] = \mu_Y = 950\mathbf{E}[] = \mu_Y = 950\mathbf{E}[] = \mu_Y = 950$$

Yes, $\bar{Y}$ is unbiased because $E[\bar{Y}] = \mu_Y$. On average across many random samples, the sample mean equals the true population mean.

(b) **Standard error:**

$$\sigma_{\bar{Y}}\sigma_{\bar{Y}}\sigma_{\bar{Y}}\sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}} = \frac{300}{\sqrt{225}} = \frac{300}{15} = \mathbf{20}$$

(c) **Sampling distribution:**

By the CLT:

$$\overset{a}{\sim} N(950,\ 400)\ \overset{a}{\sim} N(950,\ 400)\ \overset{a}{\sim} N(950,\ 400)\ \overset{a}{\sim} N(950,\ 400)$$

where the variance of $\bar{Y}$ is $\sigma^2_{\bar{Y}} = 20^2 = 400$. Equivalently, $\bar{Y} \overset{a}{\sim} N(950, 20^2)$.

(d) **Calculate $\Pr(\bar{Y} > 960)$:**

Standardize:

$$\Pr(\bar{Y} > 960) = \Pr\left(Z > \frac{960 - 950}{20}\right) = \Pr(Z > 0.50) = 1 - \Phi(0.50)$$
$$= 1 - 0.691 = \mathbf{0.309}$$

There is approximately a 30.9% chance that the sample mean exceeds 960.

(e) **Effect of increasing n to 900:**

With $n = 900$: $\sigma_{\bar{Y}} = 300/\sqrt{900} = 300/30 = 10$.

The standard error is *smaller*, so the sampling distribution of $\bar{Y}$ is more tightly concentrated around $\mu_Y = 950$. This means $\bar{Y}$ is less likely to be far from 950, so $\Pr(\bar{Y} > 960)$ **would be smaller.**

Intuitively, a larger sample gives a more precise estimate, making it less likely to observe a sample mean that deviates substantially from the population mean.

Problem 3. Conditional Expectation and the Law of Iterated Expectations (15 points)

Given: $\Pr(X = 0) = 0.70$, $\Pr(X = 1) = 0.30$.

(a) **Conditional expected values:**

For older drivers ($X = 0$):

$$\mathbf{E[Y} \mid X = 0]\mathbf{E[Y} \mid X = 0]\mathbf{E[Y} \mid X = 0]\mathbf{E[Y} \mid X = 0] = 0(0.80) + 1(0.15) + 2(0.05) = 0 + 0.15 + 0.10 = \mathbf{0.25}$$

For younger drivers ($X = 1$):

$$\mathbf{E[Y} \mid X = 1]\mathbf{E[Y} \mid X = 1]\mathbf{E[Y} \mid X = 1]\mathbf{E[Y} \mid X = 1] = 0(0.50) + 1(0.35) + 2(0.15) = 0 + 0.35 + 0.30 = \mathbf{0.65}$$

Interpretation: On average, drivers aged 25 and older have 0.25 accidents per year, while drivers under 25 have 0.65 accidents per year. Younger drivers have more than twice the expected number of accidents.

(b) **Law of Iterated Expectations:**

$$\begin{aligned}\mathbf{E[Y]} &= E[Y \mid X = 0] \cdot \Pr(X = 0) + E[Y \mid X = 1] \cdot \Pr(X = 1) \\ &= 0.25 \times 0.70 + 0.65 \times 0.30 \\ &= 0.175 + 0.195 = \mathbf{0.37}\end{aligned}$$

The overall expected number of accidents per year across all drivers is 0.37.

(c) **Var**$(Y \mid X = 1)$**:**

First, we already found $E[Y \mid X = 1] = 0.65$.

$$= (0 - 0.65)^2(0.50) + (1 - 0.65)^2(0.35) + (2 - 0.65)^2(0.15)$$

$= (0.4225)(0.50) + (0.1225)(0.35) + (1.8225)(0.15)$
$= 0.2113 + 0.0429 + 0.2734$
$= \mathbf{0.5275}$
$\mathbf{Var(Y} \mid X = 1)\mathbf{Var(Y} \mid X = 1)\mathbf{Var(Y} \mid X = 1)\mathbf{Var(Y} \mid X = 1) = \sum_y (y - 0.65)^2 \cdot \Pr(Y = y \mid X = 1) = (0 - 0.65)^2(0.50) + (1 - 0.65)^2(0.35) + (2 - 0.65)$

Causation:

No, we cannot conclude that being under 25 *causes* more accidents. This is observational data showing a *correlation* between age group and accidents, but correlation does not imply causation.

Other factors that could explain the difference:

- **Driving experience:** Younger drivers have fewer years behind the wheel.

- **Miles driven:** Younger drivers may drive more frequently or in riskier conditions (late-night driving).
- **Type of vehicle:** Younger drivers may drive older, less safe cars.
- **Risk-taking behavior:** Younger people may engage in more risky behaviors generally (omitted variable).

To establish causation, we would need a research design that isolates the effect of age—such as a regression discontinuity around an age cutoff or a controlled experiment—not just a comparison of conditional means.