Handout 2: Covariance & Correlation

ANSWER KEY

EC 282: Introduction to Econometrics

Spring 2026

1 Setup

A career center surveys recent graduates and classifies them by two random variables:

- X = number of internships completed during college (1, 2, or 3)
- Y = starting salary in thousands of dollars (60, 80, or 100)

The **joint distribution** of X and Y is given below.

	$Y = 60$	$Y = 80$	$Y = 100$	Marginal of X
$X = 1$	0.15	0.10	0.05	**0.30**
$X = 2$	0.05	0.25	0.10	**0.40**
$X = 3$	0.05	0.05	0.20	**0.30**
Marginal of Y	**0.25**	**0.40**	**0.35**	**1.00**

2 Marginal Distributions

Question 2.1: Compute the marginal distributions of X and Y by filling in the row and column totals in the table above.

Sum across each row to get the marginal distribution of X:

$$\Pr(X = 1) = 0.15 + 0.10 + 0.05 = \boxed{0.30}$$
$$\Pr(X = 2) = 0.05 + 0.25 + 0.10 = \boxed{0.40}$$
$$\Pr(X = 3) = 0.05 + 0.05 + 0.20 = \boxed{0.30}$$

Sum down each column to get the marginal distribution of Y:

$$\Pr(Y = 60) = 0.15 + 0.05 + 0.05 = \boxed{0.25}$$
$$\Pr(Y = 80) = 0.10 + 0.25 + 0.05 = \boxed{0.40}$$
$$\Pr(Y = 100) = 0.05 + 0.10 + 0.20 = \boxed{0.35}$$

✓ All marginals sum to 1.

3 Expected Values

Question 2.2: Compute $E[X]$ and $E[Y]$ using the marginal distributions:

$$E[X] = \mu_X = \sum_i x_i \cdot \Pr(X = x_i) \qquad E[Y] = \mu_Y = \sum_j y_j \cdot \Pr(Y = y_j)$$

$$\begin{aligned} E[X] = \mu_X &= 1(0.30) + 2(0.40) + 3(0.30) \\ &= 0.30 + 0.80 + 0.90 = \boxed{2.0} \end{aligned}$$

$$\begin{aligned} E[Y] = \mu_Y &= 60(0.25) + 80(0.40) + 100(0.35) \\ &= 15 + 32 + 35 = \boxed{82} \end{aligned}$$

Question 2.3: Interpret $E[X]$ and $E[Y]$ in plain language. What do these numbers tell us about the typical graduate?

On average, a graduate in this population completed **2 internships** during college and earns a starting salary of **$82,000**. These are the long-run averages—the values we would expect if we picked a graduate at random.

4 Variance and Standard Deviation

Question 2.4: Compute the variance of X using the definition:

$$\sigma_X^2 = E\big[(X - \mu_X)^2\big] = \sum_i (x_i - \mu_X)^2 \cdot \Pr(X = x_i)$$

$$\begin{aligned} \sigma_X^2 &= (1-2)^2(0.30) + (2-2)^2(0.40) + (3-2)^2(0.30) \\ &= (1)(0.30) + (0)(0.40) + (1)(0.30) \\ &= 0.30 + 0 + 0.30 = \boxed{0.60} \end{aligned}$$

Question 2.5: Similarly, compute σ_Y^2, and then the standard deviations σ_X and σ_Y.

$$\begin{aligned}\sigma_Y^2 &= (60-82)^2(0.25) + (80-82)^2(0.40) + (100-82)^2(0.35)\\ &= (-22)^2(0.25) + (-2)^2(0.40) + (18)^2(0.35)\\ &= 484(0.25) + 4(0.40) + 324(0.35)\\ &= 121 + 1.6 + 113.4 = \boxed{236}\end{aligned}$$

Standard deviations:

$$\sigma_X = \sqrt{0.60} \approx \boxed{0.7746}$$

$$\sigma_Y = \sqrt{236} \approx \boxed{15.3623}$$

5 Covariance

Question 2.6: Compute the covariance using the definition:

$$\sigma_{XY} = E\big[(X-\mu_X)(Y-\mu_Y)\big] = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) \cdot \Pr(X = x_i,\ Y = y_j)$$

We compute $(x_i - \mu_X)(y_j - \mu_Y) \cdot \Pr(X = x_i, Y = y_j)$ for each of the 9 cells. Recall $\mu_X = 2$ and $\mu_Y = 82$.

x_i	y_j	$\Pr(X=x_i, Y=y_j)$	$(x_i-2)(y_j-82)$	Product
1	60	0.15	$(-1)(-22) = 22$	$22 \times 0.15 = 3.30$
1	80	0.10	$(-1)(-2) = 2$	$2 \times 0.10 = 0.20$
1	100	0.05	$(-1)(18) = -18$	$-18 \times 0.05 = -0.90$
2	60	0.05	$(0)(-22) = 0$	$0 \times 0.05 = 0$
2	80	0.25	$(0)(-2) = 0$	$0 \times 0.25 = 0$
2	100	0.10	$(0)(18) = 0$	$0 \times 0.10 = 0$
3	60	0.05	$(1)(-22) = -22$	$-22 \times 0.05 = -1.10$
3	80	0.05	$(1)(-2) = -2$	$-2 \times 0.05 = -0.10$
3	100	0.20	$(1)(18) = 18$	$18 \times 0.20 = 3.60$

$$\sigma_{XY} = 3.30 + 0.20 - 0.90 + 0 + 0 + 0 - 1.10 - 0.10 + 3.60 = \boxed{5}$$

Note: All $X = 2$ rows contribute zero because $\mu_X = 2$ exactly.

Question 2.7: Interpret the sign of the covariance. Does the result make intuitive sense? What does it tell you about the relationship between internship experience and starting salary?

The covariance is **positive** ($\sigma_{XY} = 5$), which means that X and Y tend to move in the same direction: graduates who completed more internships tend to earn higher starting salaries. This makes intuitive sense—internship experience builds skills and professional networks that employers value.

Question 2.8: Suppose salary were measured in **dollars** instead of thousands (i.e., multiply each Y value by 1,000). How would σ_{XY} change? What does this tell you about using covariance to compare the strength of relationships?

If $Y^* = 1000 \cdot Y$, then:

$$\text{Cov}(X, Y^*) = \text{Cov}(X,\ 1000Y) = 1000 \cdot \sigma_{XY} = 1000 \times 5 = 5{,}000$$

The covariance increases by a factor of 1,000—even though the underlying relationship hasn't changed at all. This shows that **covariance depends on the units of measurement**, which makes it unreliable for comparing the strength of relationships across different scales.

6 Correlation

Question 2.9: Compute the correlation coefficient:

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$\rho_{XY} = \frac{5}{0.7746 \times 15.3623} = \frac{5}{11.90} \approx \boxed{0.42}$$

Question 2.10: Interpret ρ_{XY}. What does its value (between -1 and $+1$) tell you about the *strength* and *direction* of the linear relationship?

The correlation is approximately **0.42**, indicating a **moderate positive** linear relationship. The positive sign confirms that more internships are associated with higher salaries. The magnitude (closer to 0 than to 1) suggests that while the relationship is clearly positive, internship count alone does not perfectly predict starting salary—other factors matter too.

Question 2.11: Would ρ_{XY} change if salary were measured in dollars instead of thousands? Why is this property important?

No, ρ_{XY} would not change. If $Y^* = 1000Y$, the factor of 1,000 appears in both the numerator (through σ_{XY^*}) and the denominator (through σ_{Y^*}) and cancels out:

$$\rho_{XY^*} = \frac{1000 \cdot \sigma_{XY}}{\sigma_X \cdot 1000 \cdot \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \rho_{XY}$$

This **unitless** property is exactly why correlation is preferred over covariance for measuring the strength of a linear relationship.

7 Excel Verification

Question 2.12: Verify your calculations by entering the joint probability table in Excel and using `SUMPRODUCT()` to compute each quantity. Do your manual calculations match?

In Excel, set up the 9 outcomes with columns for x, y, and $\Pr(X=x, Y=y)$. Then:

- $E[X]$: `=SUMPRODUCT(x_col, p_col)` $= 2.0$ ✓
- $E[Y]$: `=SUMPRODUCT(y_col, p_col)` $= 82$ ✓
- σ_X^2: `=SUMPRODUCT((x_col - 2)^2, p_col)` $= 0.60$ ✓
- σ_Y^2: `=SUMPRODUCT((y_col - 82)^2, p_col)` $= 236$ ✓
- σ_{XY}: `=SUMPRODUCT((x_col - 2)*(y_col - 82), p_col)` $= 5$ ✓

All manual calculations should match.

8 Thinking Deeper

Question 2.13: We found a positive correlation between internships and starting salary. Can we conclude that doing more internships *causes* a higher starting salary? What other factors might explain this relationship?

No. Correlation does not imply causation. Several confounding factors could drive both variables:

- **Motivation/ability**: more driven students may both seek more internships *and* perform better in the job market
- **Major/field**: certain fields (e.g., finance, tech) may offer more internship opportunities *and* pay higher starting salaries
- **Family connections**: students with professional networks may secure more internships *and* higher-paying jobs
- **University prestige**: students at elite universities may have more internship access *and* better job prospects

To establish causation, we would need a research design that isolates the causal effect of internships from these confounders.

Question 2.14: A news article reports: "People who eat breakfast every day earn 20% more than people who skip breakfast." What is wrong with concluding that eating breakfast causes higher earnings?

This is a classic example of **confounding**. Breakfast eating and income may both be driven by a third variable. For example:

- People with stable, structured lifestyles may both eat breakfast regularly and hold higher-paying jobs
- Higher-income individuals may have more time and resources for morning routines
- Health-conscious behavior (correlated with breakfast eating) may also correlate with career discipline

The observed correlation between breakfast and earnings does not mean that forcing someone to eat breakfast would raise their income. This illustrates why econometrics focuses on distinguishing *correlation* from *causation*.