Econometrics 120A: Discussion Section

Midterm 1 Solutions

Anita Sammarini & Linghui Wu

Department of Economics

Covariance and correlation

Covariance and Correlation

Recap

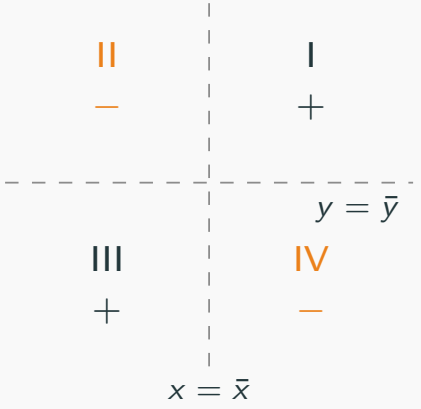
- **Sample covariance:**

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

- Each point adds one product $(x_i - \bar{x})(y_i - \bar{y})$. Its sign depends on the quadrant around $(\bar{x}, \bar{y})$.
- **Units:** units of x times units of y.
- **Correlation:**

$$r_{xy} = \frac{s_{xy}}{s_x\, s_y}$$

The units cancel, and r_{xy} is always between -1 and 1.



Sign of $(x_i - \bar{x})(y_i - \bar{y})$ in each quadrant. In I, both factors are $+$; in III, both are $-$; in II and IV, one $+$ and one $-$.

Covariance Questions

Version A, Question 1

Suppose a researcher plots a dataset of $n = 50$ store locations, where x_i is monthly advertising expenditure (in thousands of dollars) and y_i is monthly sales volume (in thousands of units).

(a) (2 pts) If the scatter plot shows data points broadly distributed evenly across all four quadrants defined by the sample means $(\bar{x}, \bar{y})$, what does this imply about the magnitude of the sample covariance s_{xy}?

(b) (2 pts) Mathematically explain how the terms $(x_i - \bar{x})(y_i - \bar{y})$ contribute to the sample covariance sum across the four quadrants relative to $(\bar{x}, \bar{y})$.

Covariance Questions

Version A, Question 1

Suppose a researcher plots a dataset of $n = 50$ store locations, where x_i is monthly advertising expenditure (in thousands of dollars) and y_i is monthly sales volume (in thousands of units).

(a) (2 pts) If the scatter plot shows data points broadly distributed evenly across all four quadrants defined by the sample means $(\bar{x}, \bar{y})$, what does this imply about the magnitude of the sample covariance s_{xy}?

Covariance Questions

Version A, Question 1

Suppose a researcher plots a dataset of $n = 50$ store locations, where x_i is monthly advertising expenditure (in thousands of dollars) and y_i is monthly sales volume (in thousands of units).

(b) (2 pts) Mathematically explain how the terms $(x_i - \bar{x})(y_i - \bar{y})$ contribute to the sample covariance sum across the four quadrants relative to $(\bar{x}, \bar{y})$.

Covariance and Correlation Questions

Version B, Question 1

(a) (1 pt) Consider two continuous variables measured on a sample of n individuals: height X (measured in meters, m) and weight Y (measured in kilograms, kg). Explain what a positive sample covariance ($s_{xy} > 0$) versus a negative sample covariance ($s_{xy} < 0$) reveals about the directional relationship between height and weight.

(b) (1 pt) What are the exact units of measurement for the sample covariance s_{xy} in this scenario?

(c) (2 pts) Why is sample covariance alone insufficient for comparing the strength of linear relationships across different pairs of variables (e.g., height vs. weight versus temperature vs. energy consumption)? How does the correlation coefficient (r_{xy}) resolve this limitation?

Covariance and Correlation Questions

Version B, Question 1

(a) (1 pt) Consider two continuous variables measured on a sample of n individuals: height X (measured in meters, m) and weight Y (measured in kilograms, kg). Explain what a positive sample covariance ($s_{xy} > 0$) versus a negative sample covariance ($s_{xy} < 0$) reveals about the directional relationship between height and weight.

Covariance and Correlation Questions

Version B, Question 1

(a) (1 pt) Consider two continuous variables measured on a sample of n individuals: height X (measured in meters, m) and weight Y (measured in kilograms, kg). Explain what a positive sample covariance ($s_{xy} > 0$) versus a negative sample covariance ($s_{xy} < 0$) reveals about the directional relationship between height and weight.

(b) (1 pt) What are the exact units of measurement for the sample covariance s_{xy} in this scenario?

Covariance and Correlation Questions

Version B, Question 1

(a) (1 pt) Consider two continuous variables measured on a sample of n individuals: height X (measured in meters, m) and weight Y (measured in kilograms, kg). Explain what a positive sample covariance ($s_{xy} > 0$) versus a negative sample covariance ($s_{xy} < 0$) reveals about the directional relationship between height and weight.

(c) (2 pts) Why is sample covariance alone insufficient for comparing the strength of linear relationships across different pairs of variables (e.g., height vs. weight versus temperature vs. energy consumption)? How does the correlation coefficient (r_{xy}) resolve this limitation?

Binomial

Binomial Random Variable

Recap: the binomial table

If $S \sim \text{Binomial}(n, p)$, the table gives $P(S = s)$: ❶ block of your n, ❷ column of your p, ❸ row of your s.

n	s	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
4	0	0.6561	0.4096	0.2401	0.1296	0.0625	0.0256	0.0081	0.0016	0.0001
4	1	0.2916	0.4096	0.4116	0.3456	0.2500	0.1536	0.0756	0.0256	0.0036
4	2	0.0486	0.1536	0.2646	0.3456	0.3750	0.3456	0.2646	0.1536	0.0486
4	3	0.0036	0.0256	0.0756	0.1536	0.2500	0.3456	0.4116	0.4096	0.2916
4	4	0.0001	0.0016	0.0081	**0.0256**	0.0625	0.1296	0.2401	0.4096	0.6561
8	5	0.0004	0.0092	0.0467	0.1239	0.2188	0.2787	0.2541	0.1468	0.0331
8	6	0.0000	0.0011	0.0100	0.0413	0.1094	0.2090	0.2965	0.2936	**0.1488**
8	7	0.0000	0.0001	0.0012	0.0079	0.0313	0.0896	0.1977	0.3355	**0.3826**
8	8	0.0000	0.0000	0.0001	0.0007	0.0039	0.0168	0.0576	0.1678	**0.4305**

The table only gives $P(S = s)$. For "at least" or "at most", add rows or use the complement: $P(S \leq 5) = 1 - \big[P(S = 6) + P(S = 7) + P(S = 8)\big]$.

Binomial Questions

Version A, Question 3

A standard medical treatment for a specific skin condition is known to have a 40% recovery rate within one week. A dermatologist tries a newly formatted ointment on a group of 4 independent patients, and all 4 patients successfully recover within the first week. The pharmaceutical representative says this completely confirms the new formula is vastly superior to the 40% standard baseline.

(a) (2 pts) What is the probability that an outcome this successful (or better) might occur purely by chance under the baseline 40% rate?

(b) (2 pts) Based on the probability in (a), do you believe the representative's claim that this proves the new ointment is superior?

Binomial Questions

Version A, Question 3

A standard medical treatment for a specific skin condition is known to have a 40% recovery rate within one week. A dermatologist tries a newly formatted ointment on a group of 4 independent patients, and all 4 patients successfully recover within the first week. The pharmaceutical representative says this completely confirms the new formula is vastly superior to the 40% standard baseline.

(a) (2 pts) What is the probability that an outcome this successful (or better) might occur purely by chance under the baseline 40% rate?

Binomial Questions

Version A, Question 3

A standard medical treatment for a specific skin condition is known to have a 40% recovery rate within one week. A dermatologist tries a newly formatted ointment on a group of 4 independent patients, and all 4 patients successfully recover within the first week. The pharmaceutical representative says this completely confirms the new formula is vastly superior to the 40% standard baseline.

(b) (2 pts) Based on the probability in (a), do you believe the representative's claim that this proves the new ointment is superior?

Binomial Questions

Version B, Question 3

A tech company asserts that its new automated software testing tool has a 90% success rate in detecting critical bugs. A quality assurance team deploys the tool on 8 independently developed code modules containing known critical bugs. Surprisingly, the tool fails to detect a bug in 3 of these modules (meaning it only successfully detected bugs in 5 modules). The developers maintain that the tool's 90% claim is still valid.

(a) (2 pts) What is the probability that such an outcome (failing to detect 3 or more bugs) might occur?

(b) (2 pts) Based on the probability in (a), do you believe that the company's claim is still correct?

Binomial Questions

Version B, Question 3

A tech company asserts that its new automated software testing tool has a 90% success rate in detecting critical bugs. A quality assurance team deploys the tool on 8 independently developed code modules containing known critical bugs. Surprisingly, the tool fails to detect a bug in 3 of these modules (meaning it only successfully detected bugs in 5 modules). The developers maintain that the tool's 90% claim is still valid.

(a) (2 pts) What is the probability that such an outcome (failing to detect 3 or more bugs) might occur?

Binomial Questions

Version B, Question 3

A tech company asserts that its new automated software testing tool has a 90% success rate in detecting critical bugs. A quality assurance team deploys the tool on 8 independently developed code modules containing known critical bugs. Surprisingly, the tool fails to detect a bug in 3 of these modules (meaning it only successfully detected bugs in 5 modules). The developers maintain that the tool's 90% claim is still valid.

(b) (2 pts) Based on the probability in (a), do you believe that the company's claim is still correct?

Normal

Normal Random Variable

Recap: two kinds of normal questions

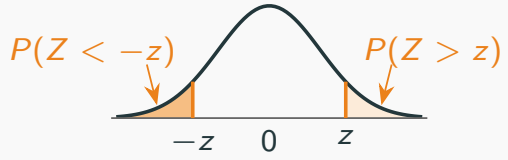
$X \sim N(\mu, \sigma^2)$
$Z = (X - \mu)/\sigma \sim N(0, 1)$
Table: $P(Z > z)$, only for $z \geq 0$

1. Find x from a probability $P(X > x)$
   - ⇨ "Top 8%" $\rightarrow P(X > x) = 0.08$
   - ⇨ Find 0.08 inside the table, read the closest z
   - ⇨ Then $x = \mu + z\,\sigma$
2. Find $P(X < x)$ given x
   - ⇨ If $z = (x - \mu)/\sigma$ is positive: $1 - P(Z > z)$
   - ⇨ If $z = (x - \mu)/\sigma$ is negative, flip it: $P(Z < -z) = P(Z > z)$

z	0.00	0.01	0.02	0.03	0.04
1.0	0.1587	0.1562	0.1539	**0.1515**	**0.1492**
1.1	0.1357	0.1335	0.1314	0.1292	0.1271
1.2	0.1151	0.1131	0.1112	0.1093	0.1075
1.3	0.0968	0.0951	0.0934	0.0918	0.0901
1.4	**0.0808**	**0.0793**	0.0778	0.0764	0.0749
2.0	**0.0228**	0.0222	0.0217	0.0212	0.0207
2.1	0.0179	0.0174	0.0170	0.0166	0.0162
2.2	0.0139	0.0136	0.0132	0.0129	0.0125
2.3	0.0107	0.0104	0.0102	**0.0099**	0.0096

Row: first decimal of z
Column: second decimal of z

2. Negative z: the two tails have the same area



Normal Questions

Version A, Question 4

Transit times for a specific cross-country shipping route average 74 hours. Assume that these transit times are normally distributed with a standard deviation of 6 hours.

(a) (2 pts) What is the transit time of a shipment at the top 8th percentile of slowest shipments? (i.e., if 8% of shipments took longer than yours, what was your transit time?)

(b) (2 pts) What is the proportion of shipments that arrive in less than 60 hours?

Normal Questions

Version A, Question 4

Transit times for a specific cross-country shipping route average 74 hours. Assume that these transit times are normally distributed with a standard deviation of 6 hours.

(a) (2 pts) What is the transit time of a shipment at the top 8th percentile of slowest shipments? (i.e., if 8% of shipments took longer than yours, what was your transit time?)

Normal Questions

Version A, Question 4

Transit times for a specific cross-country shipping route average 74 hours. Assume that these transit times are normally distributed with a standard deviation of 6 hours.

(b) (2 pts) What is the proportion of shipments that arrive in less than 60 hours?

Version B, Question 4

The lifespan of a standard delivery van's brake pads averages 32,000 miles. Assume that brake pad lifespans are normally distributed with a standard deviation of 3,500 miles.

(a) (2 pts) What is the lifespan of brake pads at the top 15th percentile? (i.e., if 15% lasted longer than yours, what would their lifespan be?)

(b) (2 pts) What is the proportion of vans whose brake pads wear out in less than 25,000 miles?

Version B, Question 4

The lifespan of a standard delivery van's brake pads averages 32,000 miles. Assume that brake pad lifespans are normally distributed with a standard deviation of 3,500 miles.

(a) (2 pts) What is the lifespan of brake pads at the top 15th percentile? (i.e., if 15% lasted longer than yours, what would their lifespan be?)

Version B, Question 4

The lifespan of a standard delivery van's brake pads averages 32,000 miles. Assume that brake pad lifespans are normally distributed with a standard deviation of 3,500 miles.

(b) (2 pts) What is the proportion of vans whose brake pads wear out in less than 25,000 miles?

Continuous densities

Recap: when is $p(x)$ a valid density?

A function $p(x)$ is a valid density only if **both** rules hold:

1. **Never negative:** $p(x) \geq 0$ for every x.
2. **Total area is 1:**

$$\int_{-\infty}^{\infty} p(x)\, dx = 1$$

Checking only the area is not enough. Probabilities are areas under $p(x)$, and a probability can never be negative.

Reminder: $\int x^k\, dx = x^{k+1}/(k+1)$.

Continuous Random Variable Questions

New question

Suppose that we have a function $p(x) = x - \frac{1}{2}$ where $0 < x < 2$.

(a) Show that the area under $p(x)$ is equal to 1.

(b) Is $p(x)$ a valid density? Explain.

Continuous Random Variable Questions

New question

Suppose that we have a function $p(x) = x - \frac{1}{2}$ where $0 < x < 2$.

(a) Show that the area under $p(x)$ is equal to 1.

Continuous Random Variable Questions

New question

Suppose that we have a function $p(x) = x - \frac{1}{2}$ where $0 < x < 2$.

(b) Is $p(x)$ a valid density? Explain.