Econometrics 120A: Discussion Section

Midterm 1 Review

Linghui Wu & Lapo Bini

Department of Economics

Bernoulli, Binomial and Normal

Bernoulli Random Variable

Bernoulli Distribution

The Bernoulli distribution is a discrete probability distribution for a random variable which takes the value 1 with probability p (success) and the value 0 with probability $1 - p$ (failure).

- ⇨ **Notation:** $X \sim \text{Bernoulli}(p)$
- ⇨ **Parameter:** p (probability of success, $0 \leq p \leq 1$)
- ⇨ **Probability Mass Function (PMF):**

$$P(X = x) = p^x(1 - p)^{1-x} \quad \text{for } x \in \{0, 1\}$$

- ⇨ **Mean (Expected Value):**

$$E[X] = p$$

- ⇨ **Variance:**

$$\text{Var}(X) = p(1 - p)$$

Binomial Random Variable

Binomial Distribution

The Binomial distribution models the number of successes in a fixed number of independent Bernoulli trials, each with the same probability of success.

- **Parameters:** n (number of trials, fixed) and p (probability of success)
- **Probability Mass Function (PMF):**

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, 2, \ldots, n$$

The binomial coefficient is defined as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- **Mean (Expected Value):**

$$E[X] = np$$

- **Variance:**

$$\text{Var}(X) = np(1-p)$$

Binomial Random Variable

Binomial distribution table

The table gives $P(X = k)$ for $X \sim \text{Binomial}(n, p)$, so we do not have to compute the formula by hand: ❶ find the block of your n, ❷ go to the column of your p, ❸ go to the row of your k.

Example: $X \sim \text{Binomial}(7, 0.3)$, $P(X = 2) = 0.3177$.

n	k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
7	0	0.4783	0.2097	0.0824	0.0280	0.0078	0.0016	0.0002	0.0000	0.0000
7	1	0.3720	0.3670	0.2471	0.1306	0.0547	0.0172	0.0036	0.0004	0.0000
7	2	0.1240	0.2753	**0.3177**	0.2613	0.1641	0.0774	0.0250	0.0043	0.0002
7	3	0.0230	0.1147	0.2269	0.2903	0.2734	0.1935	0.0972	0.0287	0.0026
7	4	0.0026	0.0287	0.0972	0.1935	0.2734	0.2903	0.2269	0.1147	0.0230
7	5	0.0002	0.0043	0.0250	0.0774	0.1641	0.2613	0.3177	0.2753	0.1240
7	6	0.0000	0.0004	0.0036	0.0172	0.0547	0.1306	0.2471	0.3670	0.3720
7	7	0.0000	0.0000	0.0002	0.0016	0.0078	0.0280	0.0824	0.2097	0.4783

The table only gives $P(X = k)$: for "at most", "at least", "fewer than" we add rows, e.g. $P(X \leq 2) = P(X = 0) + P(X = 1) + P(X = 2)$ and $P(X \geq 1) = 1 - P(X = 0)$.

Normal Random Variable

Normal Distribution

The Normal distribution is a continuous probability distribution for a random variable which can take any real value. Its density is the bell curve, symmetric around the mean.

- **Notation:** $X \sim N(\mu, \sigma^2)$
- **Parameters:** μ (mean) and σ (standard deviation)
- **Probabilities are areas under the bell curve**, so $P(X < x) = P(X \leq x)$.
- **Standardization:** if $X \sim N(\mu, \sigma^2)$ then

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1), \qquad \text{so that} \qquad P(X < x) = P\left(Z < \frac{x - \mu}{\sigma}\right)$$

  Every question about X becomes a question about Z, whose probabilities are in the table.

- **Going backwards (percentiles):** if we know the probability and want the value, we find the probability inside the table, read z at the margin, and go back with $x = \mu + z\sigma$.

Normal Random Variable

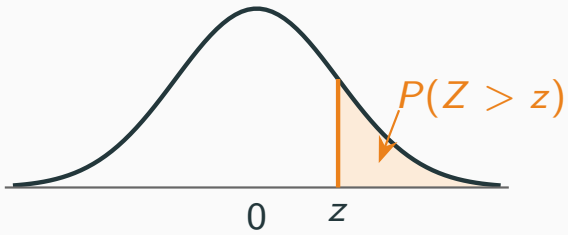
Standard normal table

The table gives $P(Z > z)$ for $z \geq 0$: ❶ the row is the first decimal of z, ❷ the column is the second decimal.

Example: $P(Z > 1.28) = 0.1003$ (row 1.2, column 0.08).

Everything else comes from the total area being 1 and the symmetry of the curve:

- ⇨ $P(Z < z) = 1 - P(Z > z)$
- ⇨ $P(Z < -z) = P(Z > z)$ (no negative z in the table)
- ⇨ $P(a < Z < b) = P(Z > a) - P(Z > b)$



z	0.06	0.07	0.08	0.09
0.9	0.1685	0.1660	0.1635	0.1611
1.0	0.1446	0.1423	0.1401	0.1379
1.1	0.1230	0.1210	0.1190	0.1170
1.2	0.1038	0.1020	**0.1003**	0.0985
1.3	0.0869	0.0853	0.0838	0.0823
1.4	0.0721	0.0708	0.0694	0.0681
1.5	0.0594	0.0582	0.0571	0.0559
1.6	0.0485	0.0475	0.0465	0.0455

Draw the bell curve and shade the area you need before touching the table: it tells you whether to use $1 - \cdot$ or the symmetry.

Bernoulli, Binomial and Normal Questions

Exercise 1: PS2, the two washers

A manufacturer needs washers between **0.118 and 0.122 inches** thick, anything outside this range is not useable. One machine shop will sell washers at **$3 per 1000 washers**. Their thickness is normally distributed with mean 0.12 and standard deviation 0.001. A second machine shop will sell washers at **$2.60 per 1000 washers**. Their thickness is normally distributed with mean 0.12 and standard deviation 0.0015 inches. Which shop offers the better deal?

Binomial Questions

Exercise 2

A coffee chain claims that **70%** of its customers pay with its loyalty app. You stand in line at one of its shops and watch the next **7 customers**: only **2** of them pay with the app. The manager says the 70% figure is still correct and you just saw an unusual group.

(a) If the claim is true, what is the probability of seeing 2 or fewer app users among 7 customers?

(b) Based on the probability in (a), do you believe the chain's claim? Explain.

Binomial Questions

Exercise 2

A coffee chain claims that **70%** of its customers pay with its loyalty app. You stand in line at one of its shops and watch the next **7 customers**: only **2** of them pay with the app. The manager says the 70% figure is still correct and you just saw an unusual group.

(a) If the claim is true, what is the probability of seeing 2 or fewer app users among 7 customers?

Binomial Questions

Exercise 2

A coffee chain claims that **70%** of its customers pay with its loyalty app. You stand in line at one of its shops and watch the next **7 customers**: only **2** of them pay with the app. The manager says the 70% figure is still correct and you just saw an unusual group.

(b) Based on the probability in (a), do you believe the chain's claim? Explain.

Discrete random variables

Discrete Probability Distribution

Discrete Distribution

⇨ **Probability Mass Function (PMF):** Defines the probability that a discrete random variable X equals a specific value.

$$P(X = x) = p(x)$$

The sum of probabilities for all possible values must equal 1:

$$\sum_{x} P(X = x) = 1$$

⇨ **Mean:**

$$E[X] = \sum_{x} x \cdot P(X = x)$$

⇨ **Variance:**

$$\text{Var}(X) = E[(X - E[X])^2] = \sum_{x} (x - E[X])^2 \cdot P(X = x)$$

⇨ **Standard deviation:**

$$\sigma = \sqrt{\text{Var}(X)}$$

Random variables and distribution questions

Exercise 3

A freelance designer submits a bid for a contract. She may win the **full contract**, worth \$8,000, with probability 0.25; a **reduced contract**, worth \$3,000, with probability 0.35; or **no contract** at all.

(a) Define a random variable X that measures the amount she earns from the bid and write down its probability distribution.

(b) Calculate the expected amount she will earn.

(c) Calculate the variance and the standard deviation of the amount.

Random variables and distribution questions

Exercise 3

A freelance designer submits a bid for a contract. She may win the **full contract**, worth \$8,000, with probability 0.25; a **reduced contract**, worth \$3,000, with probability 0.35; or **no contract** at all.

(a) Define a random variable X that measures the amount she earns from the bid and write down its probability distribution.

Random variables and distribution questions

Exercise 3

A freelance designer submits a bid for a contract. She may win the **full contract**, worth \$8,000, with probability 0.25; a **reduced contract**, worth \$3,000, with probability 0.35; or **no contract** at all.

(b) Calculate the expected amount she will earn.

Exercise 3

A freelance designer submits a bid for a contract. She may win the **full contract**, worth \$8,000, with probability 0.25; a **reduced contract**, worth \$3,000, with probability 0.35; or **no contract** at all.

(c) Calculate the variance and the standard deviation of the amount.

Continuous random variables

Continuous Random Variable

Continuous Distribution

A continuous random variable X can take any value in an interval. Its distribution is described by a **density** $p(x)$: probabilities are **areas** under the density, so the sums of the discrete case become integrals.

- ⇨ **Valid density:** $p(x) \geq 0$ for every x, and the total area must equal 1:

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

- ⇨ **Probability of an interval:** $P(a < X < b) = \int_a^b p(x)\,dx$
- ⇨ **Mean:**

$$E[X] = \int_{-\infty}^{\infty} x \cdot p(x)\,dx$$

- ⇨ **Variance:**

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x - E[X])^2 \cdot p(x)\,dx$$

The integrals run only where $p(x) > 0$. **Reminders:** $\int x^k dx = x^{k+1}/(k+1)$, $\int e^{-ax} dx = -e^{-ax}/a$, $\int u\,dv = uv - \int v\,du$.

Continuous Random Variable Questions

Exercise 4

Suppose that we have a density for a random variable $p(x) = 2e^{-2x}$ where $x > 0$.

(a) Show that this is indeed a valid density.

(b) What is the mean of this random variable?

Continuous Random Variable Questions

Exercise 4

Suppose that we have a density for a random variable $p(x) = 2e^{-2x}$ where $x > 0$.

(a) Show that this is indeed a valid density.

Continuous Random Variable Questions

Exercise 4

Suppose that we have a density for a random variable $p(x) = 2e^{-2x}$ where $x > 0$.

(b) What is the mean of this random variable?

Descriptive statistics

Measures of Central Tendencies

- **Sample Mean**: Given a sample made of n observations $\{X_1, \ldots, X_n\}$, the sample mean is computed as:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

If instead of single observations we have the frequency ω_i of each value:

$$\bar{X} = \sum_{i=1}^{F} \omega_i X_i$$

- **Sample Median**: this is the very middle observation in the data, separating the higher half from the lower half of a data sample.
- **Sample Mode**: this is the most frequently observed value in the sample.

Variance, Standard Deviation & Covariance

Definitions

⇨ **Variance**: it is a measure of dispersion, meaning it is a measure of how far a set of numbers is spread out from their average value. Given the sample $\{X_1, \ldots, X_n\}$, it is:

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

If instead of single observations we have the frequency, it becomes:

$$S_X^2 = \sum_{i=1}^{F} \omega_i (X_i - \bar{X})^2$$

⇨ **Standard deviation**: positive square root of the variance $S_X = \sqrt{S_X^2}$.

⇨ **Covariance**: it is a measure of the relationship between two variables and to what extent they change together.

$$S_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$$

Exercise 5

The hourly wages (in \$) of the 9 part-time workers of a café are:

Worker	1	2	3	4	5	6	7	8	9
Wage	12	15	15	18	20	22	25	30	60

(a) Compute the mean, the median and the mode of the wages.

(b) The owner says that "a typical worker here earns \$24 an hour". Is this a fair description? Which number would you report instead, and why?

A survey asks 200 commuters how long they take to get to work:

Minutes	10	20	30
Proportion	0.20	0.50	0.30

(c) What are the average commuting time and its standard deviation?

Exercise 5

The hourly wages (in \$) of the 9 part-time workers of a café are:

Worker	1	2	3	4	5	6	7	8	9
Wage	12	15	15	18	20	22	25	30	60

(a) Compute the mean, the median and the mode of the wages.

Exercise 5

The hourly wages (in \$) of the 9 part-time workers of a café are:

Worker	1	2	3	4	5	6	7	8	9
Wage	12	15	15	18	20	22	25	30	60

(b) The owner says that "a typical worker here earns \$24 an hour". Is this a fair description? Which number would you report instead, and why?

Descriptive Statistics using Frequencies

Exercise 5

A survey asks 200 commuters how long they take to get to work:

Minutes	10	20	30
Proportion	0.20	0.50	0.30

(c) What are the average commuting time and its standard deviation?

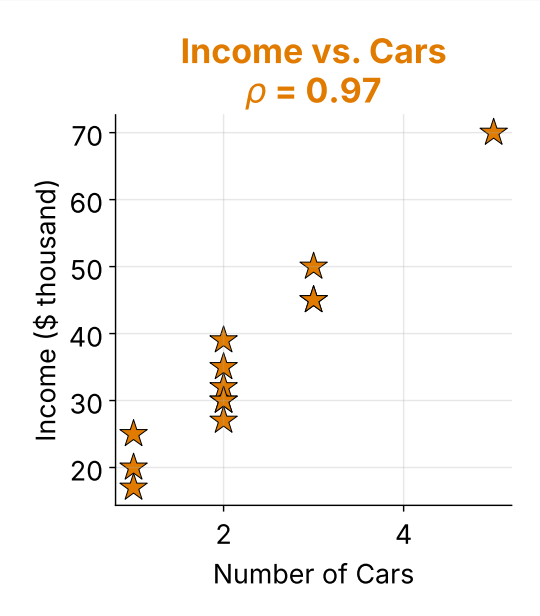
Covariance and correlation

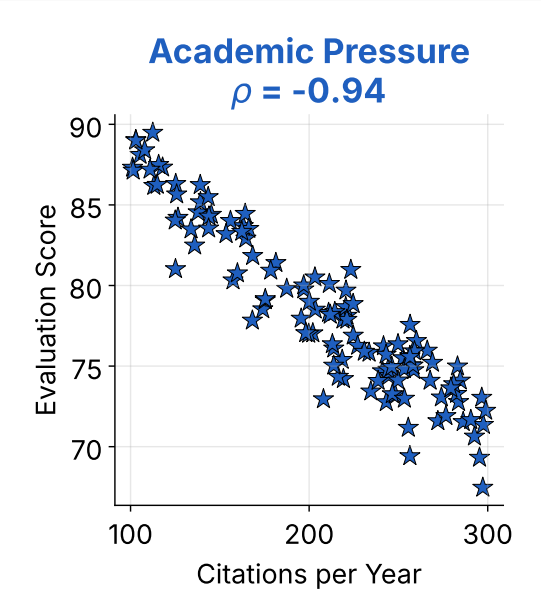
Correlation

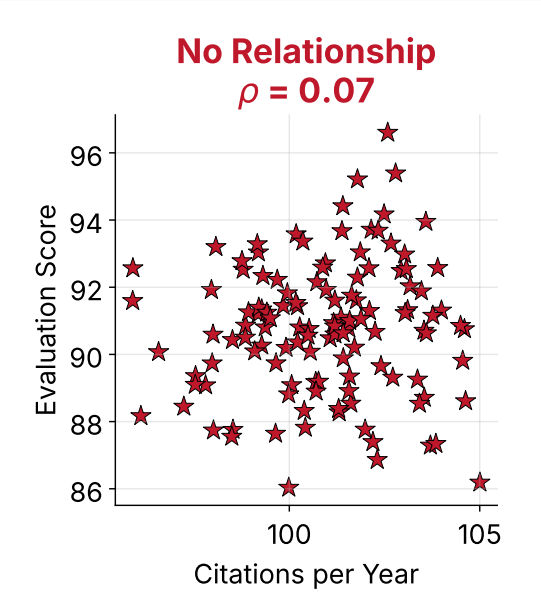
Definition

Correlation measures the extent to which two variables are **linearly** related. It is always between -1 and 1:

$$\rho_{XY} = \frac{S_{XY}}{S_X S_Y}$$







Exercise 6

Four students report the hours they studied before a quiz (X) and their score on the quiz (Y, in points):

Student	A	B	C	D
X **(hours)**	1	2	3	4
Y **(points)**	5	4	8	7

(a) Compute the sample covariance S_{XY}.

(b) Compute the correlation ρ_{XY}.

Exercise 6

Four students report the hours they studied before a quiz (X) and their score on the quiz (Y, in points):

Student	A	B	C	D
X **(hours)**	1	2	3	4
Y **(points)**	5	4	8	7

(a) Compute the sample covariance S_{XY}.

Exercise 6

Four students report the hours they studied before a quiz (X) and their score on the quiz (Y, in points):

Student	A	B	C	D
X **(hours)**	1	2	3	4
Y **(points)**	5	4	8	7

(b) Compute the correlation ρ_{XY}.

Answer key (TA only)

Answer key

Exercise 1 (the two washers)

Step 1, Normal: a washer is usable if $0.118 < X < 0.122$. Shop 1: $X \sim N(0.12, 0.001^2)$, so $z = \pm 0.002/0.001 = \pm 2$ and $P(\text{usable}) = P(-2 < Z < 2) = 1 - 2\,P(Z > 2) = 1 - 2(0.0228) = 0.9544$. Shop 2: $X \sim N(0.12, 0.0015^2)$, so $z = \pm 0.002/0.0015 = \pm 1.33$ and $P(\text{usable}) = 1 - 2\,P(Z > 1.33) = 1 - 2(0.0918) = 0.8164$.

Step 2, Bernoulli: each washer is a trial with $X_i = 1$ if usable: $p_1 = 0.9544$ for shop 1, $p_2 = 0.8164$ for shop 2.

Step 3, Binomial: the number of usable washers in a batch of 1000 is $S \sim \text{Binomial}(1000, p)$, with $E[S] = np$: shop 1 gives 954.4 usable washers for \$3, shop 2 gives 816.4 for \$2.60.

Cost per usable washer: shop 1: $3/954.4 = \$0.00314$ (\$3.14 per 1000 usable); shop 2: $2.60/816.4 = \$0.00318$ (\$3.18 per 1000 usable). **Shop 1 is the (slightly) better deal**: dearer per washer sold, but more precise, so fewer washers are thrown away. With the exact normal probabilities (0.9545 and 0.8176) the costs are \$3.14 and \$3.18: same conclusion.

Answer key

Exercises 2 to 4

2(a) Let X = number of app users among the 7 customers. Under the claim $X \sim \text{Binomial}(7, 0.7)$ and, from the table,
$P(X \le 2) = P(X=0) + P(X=1) + P(X=2) = 0.0002 + 0.0036 + 0.0250 = 0.0288$. **2(b)** If the claim were true, a group this bad or worse would show up less than 3 times in 100. Either we saw something rare or the 70% figure is too high; the evidence goes against the claim. It does not prove the share is any particular number: 7 customers is a small sample and 2.9% is not 0.

3(a) X = dollars earned from the bid: $P(X = 8{,}000) = 0.25$, $P(X = 3{,}000) = 0.35$, $P(X = 0) = 0.40$ (sums to 1). **3(b)**
$E[X] = 8{,}000(0.25) + 3{,}000(0.35) + 0(0.40) = 2{,}000 + 1{,}050 = \$3{,}050$. **3(c)**
$\text{Var}(X) = 0.25(8000 - 3050)^2 + 0.35(3000 - 3050)^2 + 0.40(0 - 3050)^2 =$
$6{,}125{,}625 + 875 + 3{,}721{,}000 = 9{,}847{,}500$ dollars^2, $\sigma = \$3{,}138$.

4(a) $p(x) = 2e^{-2x} > 0$ for $x > 0$ and $\displaystyle\int_0^\infty 2e^{-2x}\,dx = \Big[-e^{-2x}\Big]_0^\infty = 0 - (-1) = 1$. **4(b)**
$E[X] = \displaystyle\int_0^\infty x \cdot 2e^{-2x}\,dx$. By parts with $u = x$, $dv = 2e^{-2x}dx$ ($du = dx$, $v = -e^{-2x}$):
$E[X] = \Big[-xe^{-2x}\Big]_0^\infty + \displaystyle\int_0^\infty e^{-2x}\,dx = 0 + \Big[-\tfrac{1}{2}e^{-2x}\Big]_0^\infty = \tfrac{1}{2}$ (the first term is 0 because $xe^{-2x} \to 0$ as $x \to \infty$).

Answer key

Exercises 5 and 6

5(a) $\bar{X} = 217/9 = 24.11$; sorted, the 5th of 9 values is the median $= 20$; mode $= 15$. **5(b)** Not fair: the \$60 wage pulls the mean up and 7 of 9 workers earn less than \$24. The distribution is skewed right, so report the median, \$20. **5(c)** $\bar{X} = 0.2(10) + 0.5(20) + 0.3(30) = 21$ min; $S_X^2 = 0.2(10-21)^2 + 0.5(20-21)^2 + 0.3(30-21)^2 = 24.2 + 0.5 + 24.3 = 49$, so $S_X = 7$ min (weights sum to 1: no $n-1$).

6(a) $\bar{x} = 2.5$, $\bar{y} = 6$. Deviations: A $(-1.5, -1)$, B $(-0.5, -2)$, C $(0.5, 2)$, D $(1.5, 1)$, so $S_{XY} = \frac{1}{3}(1.5 + 1 + 1 + 1.5) = 5/3 = 1.67$ (hours $\times$ points). **6(b)** $S_X^2 = \frac{1}{3}(2.25 + 0.25 + 0.25 + 2.25) = 5/3$, $S_Y^2 = \frac{1}{3}(1 + 4 + 4 + 1) = 10/3$, so $\rho_{XY} = (5/3)/\sqrt{(5/3)(10/3)} = 5/\sqrt{50} = 0.71$: a fairly strong positive linear relationship.