Econ 103: Introduction to Econometrics

Lecture 3 — Expectation, Variance & Covariance

Ryan Longmuir

UCLA

Summer Session C, 2026

Reading: Hill, Griffiths & Lim (5th ed.), Probability Primer, §P.3, P.5–P.6; Stock & Watson (4th ed.), §2.2–2.3.

Where we are

Last time (Lecture 2). A random variable is described by its whole *distribution* (pmf / pdf / cdf). That is a lot of information.

Today. Summarize a distribution with a few numbers:

- its **center** — the mean (expected value),
- its **spread** — the variance and standard deviation,
- and, for *two* variables, how they **move together** — covariance and correlation.

Why these three ideas matter

Every regression coefficient we estimate later is built from exactly these pieces. The slope of a regression line, for instance, will turn out to be $\mathrm{Cov}(x, y)/\mathrm{Var}(x)$ — so today is the toolkit for the rest of the course.

A running example: the "slips" population

We reuse the population behind the pmf from Lecture 2. Ten slips in a hat; draw one at random. Define two random variables:

- $X =$ the number printed on the slip $(1, 2, 3, 4)$;
- $Y =$ an indicator: $Y = 1$ if the slip is shaded, 0 if not.

	X				
Y	1	2	3	4	$f_Y(y)$
0	0	0.1	0.2	0.3	0.6
1	0.1	0.1	0.1	0.1	0.4
$f_X(x)$	0.1	0.2	0.3	0.4	1.0

The *joint* pmf $f_{X,Y}(x, y)$, with *marginals* in the margins.

Read it two ways

- Body: $f_{X,Y}(x, y) = \mathbb{P}(X = x,\ Y = y)$.
- Right / bottom margins: the distributions of Y and X on their own.

We will compute *every* number today from this one table.

Today's plan

1. Expected value (the mean)
2. Variance & standard deviation
3. Two variables: joint, marginal, conditional
4. Conditional expectation
5. Covariance & correlation
6. Mean & variance of linear combinations

Part 1

Expected value (the mean)

The expected value

Definition

The **expected value** (or **mean**) of a discrete random variable X is the probability-weighted average of its values:

$$\mathbb{E}(X) = \sum_x x\, f_X(x) = \mu_X.$$

- It is the long-run average of X over many repetitions of the experiment.
- μ_X is a **population parameter** — a fixed feature of the population (we use Greek letters for these; later we *estimate* them from a sample).

Heads-up on names: the "mean" can refer to this *population* mean μ_X *or* to a *sample* average $\bar{x}$. They are different objects — keep track of which one is meant.

Example: the mean of X, and the mean of an indicator

Number on the slip, X:

$$\begin{aligned}\mathbb{E}(X) &= \textstyle\sum_x x\, f_X(x) \\ &= 1(0.1) + 2(0.2) + 3(0.3) + 4(0.4) \\ &= 3.\end{aligned}$$

Draw thousands of slips and average the numbers — the average settles down to 3.

Paying off a Lecture-2 promise

For the **indicator** Y (Bernoulli):

$$\mathbb{E}(Y) = 0(1-p) + 1(p) = p.$$

The mean of a 0/1 variable is the proportion of ones. Here $\mathbb{E}(Y) = 0.4 = \mathbb{P}(\text{shaded})$.

This is why, later, a regression on an indicator reads off a group's *share* or a *treatment effect*.

The expected value of a function of X

Any function $g(X)$ of a random variable is itself random. Its mean weights the *transformed* values by the *same* probabilities:

$$\mathbb{E}[g(X)] \;=\; \sum_x g(x)\, f_X(x).$$

Example (Second moment of X)

With $g(X) = X^2$,

$$\begin{aligned}\mathbb{E}(X^2) &= \textstyle\sum_x x^2 f_X(x) \\ &= 1(0.1) + 4(0.2) + 9(0.3) + 16(0.4) \\ &= 10.\end{aligned}$$

A trap to avoid

In general

$$\mathbb{E}[g(X)] \;\neq\; g(\mathbb{E}(X))\,.$$

Here $\mathbb{E}(X^2) = 10$ but $\big(\mathbb{E}X\big)^2 = 3^2 = 9$. We will use $\mathbb{E}(X^2)$ in a moment to get the variance.

Rules for expected values

Let a, b, c be constants and X, Y random variables. Expectation is a linear operator:

Property (Linearity of expectation)

$$\begin{aligned}\mathbb{E}(aX + b) &= a\,\mathbb{E}(X) + b,\\ \mathbb{E}[g_1(X) + g_2(X)] &= \mathbb{E}[g_1(X)] + \mathbb{E}[g_2(X)],\\ \mathbb{E}(aX + bY + c) &= a\,\mathbb{E}(X) + b\,\mathbb{E}(Y) + c.\end{aligned}$$

In words. "The expected value of a sum is the sum of the expected values," and constants pass straight through.

One caution about products

Linearity is about *sums*. For *products*, $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$ holds only when X and Y are independent — otherwise the covariance (later today) gets in the way.

Part 2

Variance & standard deviation

Measuring spread: the variance

Definition

The **variance** of X is the expected squared distance from the mean:

$$\mathrm{Var}(X) \;=\; \mathbb{E}\big[(X-\mu_X)^2\big] \;=\; \sigma_X^2.$$

The **standard deviation** $\sigma_X = \sqrt{\mathrm{Var}(X)}$ is in the same units as X.

A larger variance means the distribution is more spread out about its mean.

The computational formula (use this one)

$$\mathrm{Var}(X) \;=\; \mathbb{E}(X^2) - \mu_X^2.$$

Derivation: expand $\mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - 2\mu\mathbb{E}(X) + \mu^2 = \mathbb{E}(X^2) - \mu^2$, since $\mathbb{E}(X) = \mu$.

Example: variance of X and of an indicator

Number on the slip, X: we found $\mathbb{E}(X) = 3$ and $\mathbb{E}(X^2) = 10$, so

$$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu_X^2 = 10 - 3^2 = 1,$$

and $\sigma_X = \sqrt{1} = 1$.

Variance of a Bernoulli

For the indicator Y with $\mathbb{E}(Y) = p$ (and $Y^2 = Y$, so $\mathbb{E}(Y^2) = p$):

$$\mathrm{Var}(Y) = p - p^2 = p(1-p).$$

Here $\mathrm{Var}(Y) = 0.4(0.6) = 0.24$, so $\sigma_Y = \sqrt{0.24} \approx 0.49$.

A coin is most uncertain at $p = \frac{1}{2}$, where $p(1-p)$ is largest.

Variance under a linear transformation

What happens to spread when we rescale and shift? Let a, b be constants:

Property (Mean and variance of $a + bX$)

$$\mathbb{E}(a + bX) = a + b\,\mu_X, \qquad \mathrm{Var}(a + bX) = b^2\,\mathrm{Var}(X), \qquad \sigma_{a+bX} = |b|\,\sigma_X.$$

- An additive constant a shifts the whole distribution — it moves the mean but leaves the spread unchanged.
- A multiplicative constant b rescales — it multiplies the standard deviation by $|b|$ and the variance by b^2.

Example (After-tax earnings (Stock & Watson))

Tax pre-tax earnings X at 20% and add a \$2000 grant: $Y = 2000 + 0.8X$. Then $\mu_Y = 2000 + 0.8\,\mu_X$ and $\sigma_Y = 0.8\,\sigma_X$ — the spread of take-home pay is 80% that of pre-tax pay.

A useful special case: standardization

Combining the two rules, we can turn *any* X into a variable with mean 0 and variance 1. Subtract the mean and divide by the standard deviation:

$$Z = \frac{X - \mu_X}{\sigma_X}.$$

Using the linear rules with $a = -\mu_X/\sigma_X$ and $b = 1/\sigma_X$:

$$\mathbb{E}(Z) = 0, \qquad \mathrm{Var}(Z) = \frac{\mathrm{Var}(X)}{\sigma_X^2} = 1.$$

Why we care

Z is unit-free and measures "how many standard deviations from the mean." This is exactly the move behind the *Z-score* and the standard Normal table — the heart of next lecture.

Part 3

Two variables: joint, marginal, conditional

Joint and marginal distributions

Most economic questions involve *two* variables at once (income *and* education; price *and* quantity).

Joint pmf

$f_{X,Y}(x, y) = \mathbb{P}(X = x,\ Y = y)$ — the probability the two outcomes occur together. The entries sum to 1.

Marginal pmf

The distribution of one variable alone, obtained by summing the joint over the other:

$$f_X(x) = \sum_y f_{X,Y}(x, y).$$

From the slips table:

- Sum down each column $\Rightarrow$ $f_X = (0.1, 0.2, 0.3, 0.4)$.
- Sum across each row $\Rightarrow$ $f_Y = (0.6,\ 0.4)$.

e.g. $\mathbb{P}(\text{shaded}) = f_Y(1)$
$= 0.1 + 0.1 + 0.1 + 0.1 = 0.4$.

Conditional distributions

Often we want the distribution of X *within a subpopulation* fixed by Y — conditioning shrinks the population to those cases:

$$f_{X \mid Y}(x \mid y) = \mathbb{P}(X = x \mid Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}.$$

Example (Shaded slips only)

Among shaded slips ($Y = 1$, probability 0.4),

$$f_{X \mid Y}(x \mid 1) = \frac{0.1}{0.4} = 0.25$$

for each x — equally likely once we know the slip is shaded.

Rain and the commute (Stock & Watson)

$X = 0$ rain, $Y = 0$ long commute. With $\mathbb{P}(\text{rain}) = 0.30$ and a rainy-&-long probability of 0.15,

$$\mathbb{P}(\text{long} \mid \text{rain}) = \frac{0.15}{0.30} = 0.50.$$

Independence

Definition

X and Y are **independent** if knowing one tells you *nothing* about the other — equivalently, for *all* x, y,

$$f_{X \mid Y}(x \mid y) = f_X(x) \quad \Longleftrightarrow \quad f_{X,Y}(x, y) = f_X(x)\, f_Y(y).$$

(The joint factors into the product of the marginals.)

Example (The slips are *not* independent)

Check the corner $x = 1$, $y = 1$*:*

$$f_{X,Y}(1,1) = 0.1 \;\neq\; f_X(1)\, f_Y(1) = (0.1)(0.4) = 0.04.$$

A single violated cell is enough — X and Y are dependent. (Makes sense: shaded slips are never a "1".)

Part 4

Conditional expectation

Conditional expectation

The **conditional expectation** $\mathbb{E}(X \mid Y = y)$ is just the mean computed with the conditional pmf:

$$\mathbb{E}(X \mid Y = y) \;=\; \sum_x x\, f_{X \mid Y}(x \mid y).$$

It answers questions like "what is the mean wage *among* people with 16 years of education?", $\mathbb{E}(\text{WAGE} \mid \text{EDUC} = 16)$.

Example (Slips, given shaded)

$$\begin{aligned}\mathbb{E}(X \mid Y = 1) &= \textstyle\sum_x x\, f_{X \mid Y}(x \mid 1) \\ &= (1+2+3+4)(0.25) = 2.5.\end{aligned}$$

Note 2.5 is not a value X can take — an expected value need not be attainable.

It is a function of the conditioning value

Conditioning on white slips instead gives

$$\mathbb{E}(X \mid Y = 0) = \tfrac{10}{3} \approx 3.33,$$

while the *unconditional* mean is $\mathbb{E}(X) = 3$. So $\mathbb{E}(X \mid Y)$ varies with Y.

The law of iterated expectations

The conditional means must "average back" to the overall mean, weighted by how often each condition occurs.

Property (Law of iterated expectations)

$$\mathbb{E}(X) \;=\; \sum_{y} \mathbb{E}(X \mid Y = y)\, f_Y(y) \;=\; \mathbb{E}[\mathbb{E}(X \mid Y)]\,.$$

Example (Check it on the slips)

$$\mathbb{E}(X) = \underbrace{\tfrac{10}{3}}_{\mathbb{E}(X \mid Y=0)}(0.6) + \underbrace{2.5}_{\mathbb{E}(X \mid Y=1)}(0.4) = 2.0 + 1.0 = 3 \checkmark$$

Intuition (Stock & Watson): mean adult height is the mean height of men and of women, weighted by their population shares.

Conditional variance — and a preview of regression

We can also measure *spread* within a subpopulation:

$$\mathrm{Var}(X \mid Y = y) = \mathbb{E}\big[(X - \mathbb{E}(X \mid Y = y))^2 \,\big|\, Y = y\big].$$

For the slips, $\mathrm{Var}(X \mid Y = 1) = \frac{5}{4}$ while $\mathrm{Var}(X \mid Y = 0) = \frac{5}{9}$: the spread of X differs across subpopulations, and either can exceed or fall short of the unconditional $\mathrm{Var}(X) = 1$.

Why conditional expectation is the punchline of the course

Among *all* functions $g(X)$, the conditional mean $\mathbb{E}(Y \mid X)$ is the best predictor of Y from X — it minimizes the mean squared prediction error $\mathbb{E}\big[(Y - g(X))^2\big]$.
The regression line we build starting in Lecture 5 is precisely a model for $\mathbb{E}(Y \mid X)$.

Part 5

Covariance & correlation

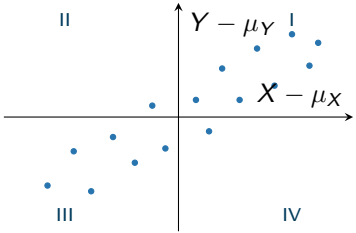
Covariance: do two variables move together?

Definition

The **covariance** of X and Y measures their linear association:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] = \mathbb{E}(XY) - \mu_X \mu_Y = \sigma_{XY}.$$

- $\sigma_{XY} > 0$: when X is above its mean, Y *tends* to be too (points in quadrants I & III).
- $\sigma_{XY} < 0$: they move in *opposite* directions (quadrants II & IV).
- $\sigma_{XY} \approx 0$: no *linear* tendency.



Positive covariance: mostly I & III.

Example: covariance of the slips

First the cross-moment (only the shaded row $Y = 1$ contributes, since $Y = 0$ kills the product):

$$\mathbb{E}(XY) = \sum_{x,y} xy\, f_{X,Y}(x, y) = (1 + 2 + 3 + 4)(1)(0.1) = 1.$$

Then, using $\mathbb{E}(X) = 3$ and $\mathbb{E}(Y) = 0.4$,

$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y = 1 - (3)(0.4) = -0.2.$$

The covariance is negative: larger numbers are relatively more common on the *white* slips, so a high X goes with $Y = 0$. Consistent with the dependence we found earlier.

Correlation: a unit-free covariance

Covariance has awkward units (here, "slip-number $\times$ shaded"), and its size is hard to read. Dividing by the standard deviations fixes both:

Definition

$$\rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)}} = \frac{\sigma_{XY}}{\sigma_X\,\sigma_Y}, \qquad -1 \le \rho_{XY} \le 1.$$

For the slips:

$$\rho_{XY} = \frac{-0.2}{\sqrt{1}\,\sqrt{0.24}} \approx -0.41.$$

$\rho = \pm 1$ exactly when X is a perfect linear function of Y; $\rho = 0$ means no linear association.

A real-data anchor

The food-expenditure vs. income data from Lecture 1 had correlation $\rho \approx 0.62$ — a moderate, *positive* linear association, matching that upward-sloping cloud.

Independence, covariance, and a crucial caveat

One direction holds...

If X and Y are independent, then $\mathrm{Cov}(X, Y) = 0$ and $\rho_{XY} = 0$.

...but the converse does *not*

$\mathrm{Cov}(X, Y) = 0$ does not imply independence. Covariance only sees *linear* association; variables can be tightly related in a *nonlinear* way yet have zero covariance.

Example (Zero covariance, total dependence)

Let points lie on the circle $X^2 + Y^2 = 1$, symmetric about the axes. Then $\mathrm{Cov}(X, Y) = 0$, yet X and Y are completely dependent — knowing X pins Y down to $\pm\sqrt{1 - X^2}$.

Part 6

Mean & variance of linear combinations

Combining random variables: the mean

We constantly build new variables as weighted sums of others (a portfolio, a sample average, a regression fit). Start with the mean — it is always linear:

Property (Mean of a linear combination)

$$\mathbb{E}(aX + bY + c) \;=\; a\,\mathbb{E}(X) + b\,\mathbb{E}(Y) + c,$$

whether or not X and Y are independent.

This extends to any number of terms:

$$\mathbb{E}\left(\sum_i a_i X_i\right) = \sum_i a_i\,\mathbb{E}(X_i).$$

No assumptions needed — expectation does not care about dependence.

Combining random variables: the variance

Variance is not linear — a covariance term appears:

Property (Variance of a linear combination)

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\,\mathrm{Cov}(X, Y).$$

Two special cases worth memorizing:

$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y),$$

$$\mathrm{Var}(X - Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) - 2\mathrm{Cov}(X, Y).$$

The headline

The variance of a sum is *not* the sum of the variances — unless the variables are uncorrelated.

The independent (or uncorrelated) case

When $\mathrm{Cov}(X, Y) = 0$ — in particular when X and Y are independent — the cross term vanishes and variance *does* add:

$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y), \qquad \mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Looking ahead to Lecture 4

The **sample mean** $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a linear combination of independent draws. These rules give

$$\mathbb{E}(\bar{X}) = \mu, \qquad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}.$$

The variance shrinks as n grows — the reason larger samples are more informative, and the seed of the Central Limit Theorem.

Recap

One variable

- Mean: $\mathbb{E}(X) = \sum_x x f_X(x)$
- $\mathbb{E}$ is linear; $\mathbb{E}[g(X)] \neq g(\mathbb{E}X)$
- Var: $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mu^2$
- $\mathrm{Var}(a + bX) = b^2\mathrm{Var}(X)$
- Indicator: $\mathbb{E} = p$, $\mathrm{Var} = p(1-p)$

Two variables

- Joint $\to$ marginal (sum out) $\to$ conditional (divide)
- Indep. $\Leftrightarrow f_{X,Y} = f_X f_Y$
- $\mathrm{Cov} = \mathbb{E}(XY) - \mu_X\mu_Y$; $\rho = \sigma_{XY}/(\sigma_X\sigma_Y)$
- Indep. $\Rightarrow \mathrm{Cov} = 0$ (not conversely)
- $\mathrm{Var}(X+Y) = \mathrm{Var}X + \mathrm{Var}Y + 2\mathrm{Cov}$

The thread to regression

$\mathbb{E}(Y \mid X)$ is the best predictor of Y; the regression slope will be $\mathrm{Cov}(X, Y)/\mathrm{Var}(X)$. **Next time (Lecture 4):** the Normal distribution, sampling, and the Central Limit Theorem.

Questions?