Variance, Covariance, and Correlation

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's Reading

Required

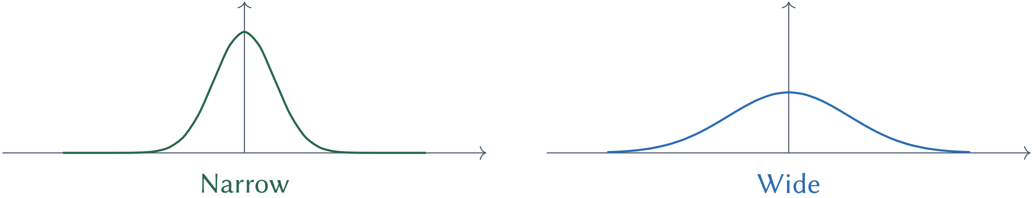
- **Aronow & Miller**, §2.1 (variance) and §2.2.1–2.2.2 (covariance, correlation)
- **Blackwell**, Ch. 2.4–2.5: Summaries of distributions

Key concepts: variance, standard deviation, covariance, correlation, independence vs. uncorrelatedness.

Beyond the Mean

Last time: Expected value $\mathbb{E}[X]$ tells us the center of a distribution.

But consider two distributions with the same mean:



Both have mean 0, but they're very different distributions.
We need a measure of spread.

Definition: Variance

Variance

The **variance** of a random variable X is:

$$\text{Var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right]$$

Interpretation:

- Average squared deviation from the mean
- How far, on average, is X from its expected value?
- Larger variance = more spread out

Also written σ^2 or σ_X^2. Always non-negative: $\text{Var}(X) \geq 0$.

The Computational Formula

Useful Alternative

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$$

Derivation:

$$\begin{aligned}
\text{Var}(X) &= \mathbb{E}[(X - \mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2 - 2X\,\mathbb{E}[X] + (\mathbb{E}[X])^2] \\
&= \mathbb{E}[X^2] - 2\,\mathbb{E}[X]\,\mathbb{E}[X] + (\mathbb{E}[X])^2 \\
&= \mathbb{E}[X^2] - (\mathbb{E}[X])^2
\end{aligned}$$

This is often easier to compute!

Remember: $\mathbb{E}[X^2] \neq (\mathbb{E}[X])^2$ unless $\text{Var}(X) = 0$.

Example: Variance of a Die Roll

Setup: Roll a fair die. Find $\text{Var}(X)$.

From Monday: $\mathbb{E}[X] = 3.5$ and $\mathbb{E}[X^2] = \frac{91}{6}$

Using the computational formula:

$$\begin{aligned}\text{Var}(X) &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \\ &= \frac{91}{6} - (3.5)^2 \\ &= \frac{91}{6} - 12.25 \\ &= 15.167 - 12.25 = 2.917\end{aligned}$$

Standard deviation: $\text{SD}(X) = \sqrt{2.917} \approx 1.71$

Example: Bernoulli Variance

Setup: $X \sim \text{Bernoulli}(p)$

We know: $\mathbb{E}[X] = p$

Find $\mathbb{E}[X^2]$: Since $X \in \{0, 1\}$, we have $X^2 = X$, so $\mathbb{E}[X^2] = \mathbb{E}[X] = p$

Variance:

$$\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = p - p^2 = p(1 - p)$$

Key Result

For $X \sim \text{Bernoulli}(p)$: $\text{Var}(X) = p(1 - p)$
Maximum variance at $p = 0.5$: most uncertainty when outcome is 50-50.

Standard Deviation

Definition

The **standard deviation** is the square root of variance:

$$\text{SD}(X) = \sigma_X = \sqrt{\text{Var}(X)}$$

Why use SD instead of variance?

- Same units as X (variance has squared units)
- More interpretable: "typical deviation from the mean"
- For normal distributions: about 68% of data within 1 SD of mean

We'll use both $\text{Var}(X)$ and $\text{SD}(X)$ throughout the course.

Properties of Variance: Basics

1. $\text{Var}(X) \geq 0$ always (squared deviations can't be negative)
2. $\text{Var}(c) = 0$ for any constant c (no spread if no randomness)
3. $\text{Var}(aX + b) = a^2 \text{Var}(X)$

Property 3 in words: Adding a constant shifts the distribution but doesn't change spread. Scaling by a multiplies the spread by $|a|$ (variance by a^2).

Variance of Sums: A Critical Difference from Expectation

For **independent** X and Y:

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

Warning: Independence is required!

Compare to expectation:

- $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$ (always)
- $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$ (only if independent)

If X and Y are dependent, we need covariance.

Why Variance Needs Independence

Counter-example: Let $Y = -X$.

Then:

$$\text{Var}(X + Y) = \text{Var}(X + (-X)) = \text{Var}(0) = 0$$

But:

$$\begin{aligned} \text{Var}(X) + \text{Var}(Y) &= \text{Var}(X) + \text{Var}(-X) \\ &= \text{Var}(X) + (-1)^2\,\text{Var}(X) \\ &= 2\,\text{Var}(X) \neq 0 \end{aligned}$$

The general formula (which we'll derive shortly):

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

Example: Binomial Variance

Setup: $X \sim \text{Binomial}(n, p)$, where $X = X_1 + \cdots + X_n$ and $X_i \overset{iid}{\sim} \text{Bernoulli}(p)$.

Since the X_i are **independent**:

$$\begin{aligned} \text{Var}(X) &= \text{Var}(X_1 + \cdots + X_n) \\ &= \text{Var}(X_1) + \cdots + \text{Var}(X_n) \\ &= p(1-p) + \cdots + p(1-p) \\ &= np(1-p) \end{aligned}$$

Key Result

For $X \sim \text{Binomial}(n, p)$: $\text{Var}(X) = np(1-p)$

Variance: Quick Reference

Distribution	Notation	$\mathbb{E}[X]$	$\text{Var}(X)$
Bernoulli	Bernoulli(p)	p	$p(1-p)$
Binomial	Binomial(n, p)	np	$np(1-p)$
Poisson	Poisson(λ)	λ	λ
Uniform	Uniform(a, b)	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	$N(\mu, \sigma^2)$	μ	σ^2
Exponential	Exp(λ)	$\frac{1}{\lambda}$	$\frac{1}{\lambda^2}$

Notice: Poisson has mean = variance. Normal's σ^2 parameter IS the variance.

From One Variable to Two

So far: Summaries for a single random variable X

- Center: $\mathbb{E}[X]$
- Spread: $\text{Var}(X)$

Now: What if we have *two* random variables X and Y?

New question: Do X and Y move together?

- When X is high, is Y also high? (positive relationship)
- When X is high, is Y low? (negative relationship)
- Is there no systematic pattern? (no relationship)

We need covariance.

Definition: Covariance

Covariance

The **covariance** of random variables X and Y is:

$$\text{Cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right]$$

Interpretation:

- Average of the product of deviations from means
- $\text{Cov}(X, Y) > 0$: X and Y tend to be above/below their means together
- $\text{Cov}(X, Y) < 0$: when one is above its mean, the other tends to be below
- $\text{Cov}(X, Y) = 0$: no linear relationship (uncorrelated)

Computational Formula for Covariance

Useful Alternative

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$$

Derivation:

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\
&= \mathbb{E}[XY - X\,\mathbb{E}[Y] - Y\,\mathbb{E}[X] + \mathbb{E}[X]\,\mathbb{E}[Y]] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[Y]\,\mathbb{E}[X] + \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]
\end{aligned}$$

Note: Variance is a special case: $\text{Var}(X) = \text{Cov}(X, X)$

Key Insight: Variance is a Special Case of Covariance

Unifying Principle

$$\text{Cov}(X, X) = \text{Var}(X)$$

This is beautiful: Variance and covariance are the same concept.

Covariance measures how two variables move together.
Variance measures how a variable moves with *itself*.

This unification comes from Aronow & Miller's "agnostic" framework.

Properties of Covariance

1. $\text{Cov}(X, Y) = \text{Cov}(Y, X)$ (symmetric)
2. $\text{Cov}(X, c) = 0$ for any constant c
3. $\text{Cov}(aX + b, cY + d) = ac \cdot \text{Cov}(X, Y)$

Bilinearity (the key property):

$$\text{Cov}(X + Y, Z) = \text{Cov}(X, Z) + \text{Cov}(Y, Z)$$

Bilinearity lets us expand sums inside covariance—crucial for deriving variance of sums.

Variance of Sums (General Formula)

Using bilinearity of covariance:

$$\begin{aligned} \text{Var}(X+Y) &= \text{Cov}(X+Y, X+Y) \\ &= \text{Cov}(X,X) + \text{Cov}(X,Y) + \text{Cov}(Y,X) + \text{Cov}(Y,Y) \\ &= \text{Var}(X) + 2\,\text{Cov}(X,Y) + \text{Var}(Y) \end{aligned}$$

General Formula

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

Special case: If $X \perp\!\!\!\perp Y$ (independent), then $\text{Cov}(X,Y) = 0$, so:

$$\text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y)$$

Independence $\Rightarrow$ Zero Covariance

Theorem

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

Proof:
If $X \perp\!\!\!\perp Y$, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$. Therefore:

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] = \mathbb{E}[X]\,\mathbb{E}[Y] - \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

Warning: The converse is FALSE!
$\text{Cov}(X, Y) = 0 \not\Rightarrow X \perp\!\!\!\perp Y$

Zero covariance means no *linear* relationship. There could still be a nonlinear one.

Zero Covariance $\nRightarrow$ Independence

Classic example: Let $X \sim \text{Uniform}(-1, 1)$ and $Y = X^2$.

Y is completely determined by X—they're totally dependent!

But check the covariance:

$$\begin{aligned}
\text{Cov}(X, Y) &= \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y] \\
&= \mathbb{E}[X \cdot X^2] - \mathbb{E}[X] \cdot \mathbb{E}[X^2] \\
&= \mathbb{E}[X^3] - 0 \cdot \mathbb{E}[X^2] \quad \text{(since } \mathbb{E}[X] = 0 \text{ by symmetry)} \\
&= \mathbb{E}[X^3] = 0 \quad \text{(by symmetry)}
\end{aligned}$$

Lesson: Cov $= 0$ only rules out *linear* relationships.

Correlation: Standardized Covariance

Problem with covariance: Units depend on X and Y.

Is $\text{Cov}(X, Y) = 1000$ big? Depends on the scales!

Correlation

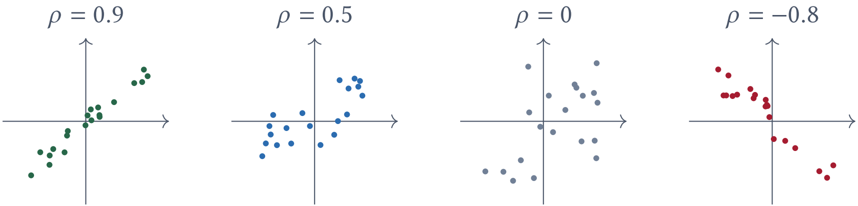
The **correlation** (or Pearson correlation coefficient) is:

$$\text{Corr}(X, Y) = \rho_{XY} = \frac{\text{Cov}(X, Y)}{\text{SD}(X) \cdot \text{SD}(Y)} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\,\text{Var}(Y)}}$$

Key property: $-1 \leq \rho \leq 1$ always

- $\rho = 1$: perfect positive linear relationship
- $\rho = -1$: perfect negative linear relationship
- $\rho = 0$: no linear relationship (uncorrelated)

Visualizing Correlation



As $|\rho| \to 1$, points cluster closer to a line.

Properties of Correlation

Key Properties

1. $-1 \leq \rho_{XY} \leq 1$
2. $\rho_{XY} = \rho_{YX}$ (symmetric)
3. $\rho_{XY} = \pm 1$ iff $Y = a + bX$ for some constants a, b
   - $\rho = 1$ if $b > 0$; $\rho = -1$ if $b < 0$
4. $\text{Corr}(aX + b, cY + d) = \text{sign}(ac) \cdot \rho_{XY}$ (if $ac \neq 0$)
5. Correlation is **unit-free** (dimensionless)

Linear transformations don't change the magnitude of correlation, only possibly the sign.

Correlation Is Not Causation

This will be a recurring theme in this course.

Example: Democracy and peace are positively correlated.

Does democracy cause peace? *Maybe*—but there are **confounders**: wealth, trade, alliances, geography all correlate with both.

Three possible explanations for $\text{Corr}(X, Y) \neq 0$:

1. X causes Y
2. Y causes X
3. Some third variable Z causes both (confounding)

Correlation describes association. Causation requires more.

Example: Education and Income

Suppose: $\text{Corr}(\text{Education}, \text{Income}) = 0.4$

What does this tell us?

- There's a positive linear association
- People with more education tend to have higher income
- The relationship is moderate (not perfect)

What does this NOT tell us?

- Whether education *causes* higher income
- Maybe ability drives both
- Maybe family background drives both
- Maybe income enables more education (reverse causality)

Distinguishing these is the hard work of causal inference.

Key Takeaways

1. **Variance** measures spread: $\text{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
2. **Covariance** is general; variance is the special case $\text{Cov}(X, X)$
3. **Independence** $\Rightarrow$ Cov $= 0$, but **not** vice versa

The big idea: Correlation measures linear association, not causation.

Next week: Joint distributions and the conditional expectation function.

For Monday

Topic: Joint Distributions and the CEF

Reading:

- A&M §1.3 and §2.2.3–2.2.4
- Blackwell Ch. 1

The CEF is the heart of this course—regression approximates it.