Covariance and Correlation

Gov 51: Data Analysis and Politics

Scott Cunningham

Harvard University

Week 3, Thursday
February 13, 2026

Announcement: Section Attendance Policy

New: Section attendance is now worth **5 bonus points**.

How it works:

- Attend at least 70% of sections → earn 5 bonus points
- Your final grade = (points earned + bonus) / 105
- Example: 85 points + 5 bonus = $90/105 \approx 85.7\%$

Conflicts?

- If you have a scheduling conflict, you can complete a makeup assignment (one-page reflection on the section content)
- Talk to George if you need this accommodation

Sections reinforce lecture material and prepare you for exams. We strongly encourage attendance.

A Puzzle from Problem Set 1

A Result You Just Computed

In PS1, you calculated two means for commute time:

```r
unweighted_mean <- mean(commuters$TRANTIME)
weighted_mean <- weighted.mean(commuters$TRANTIME, commuters
    $PERWT)

# Results:
# Unweighted: 27.22 minutes
# Weighted:   27.19 minutes
# Difference: -0.03 minutes
```

These are almost identical. Why?

But Sometimes They're Very Different

Remember from Week 2—state approval ratings:

Measure	Value
Unweighted mean (average across states)	43.3%
Weighted mean (average across people)	44.5%
Difference	**+1.2 percentage points**

Here the weighted mean is noticeably higher.

Why does weighting matter a lot in one case but not the other?

Today's Question

What determines when weighted and un-weighted means are the same vs. different?

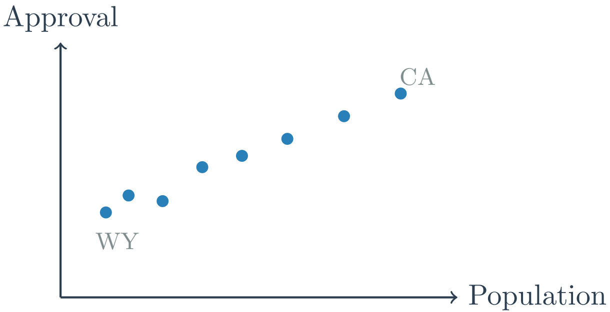
To answer this, we need a new concept: **covariance**.

By the end of today, you'll be able to predict when weighting will matter—before you even calculate anything.

Introducing Covariance

The Intuition: Do Two Things Move Together?

Question: Do states with larger populations have higher or lower approval ratings?



When California has high approval AND high population, that's meaningful.

Historical Note: Where Covariance Came From

Late 1800s: Francis Galton and Karl Pearson developed these ideas.

- Galton was studying heredity: Do tall parents have tall children?
- He noticed that variables "co-vary"—they vary *together*
- Pearson formalized the mathematics

The concept of "co-variation" became **covariance**.

We're using 130-year-old tools because they work.

Building the Formula: Step by Step

Start with something familiar: **deviations from the mean**.

For each observation:

- $x_i - \bar{x}$ = how far is this value from its mean?

Centering Property: Deviations from the mean *always* sum to zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

This is a fundamental property of the mean—it's the "balance point" of the data.

But What About Two Variables?

You might think: if $\sum(x_i - \bar{x}) = 0$, does $\sum(x_i - \bar{x})(y_i - \bar{y})$ also equal zero?

No! Let's see why with an example:

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	1	2	-2	-2	$+4$
2	3	4	0	0	0
3	5	6	$+2$	$+2$	$+4$
Sum			**0**	**0**	$\mathbf{+8}$

$(\bar{x} = 3,\ \bar{y} = 4)$

Individual deviations sum to zero, but their **products** do not!

Why the Product Doesn't Cancel Out

When two variables move **together**:

- x above mean AND y above mean $\rightarrow (+)(+) =$ positive
- x below mean AND y below mean $\rightarrow (-)(-) =$ positive

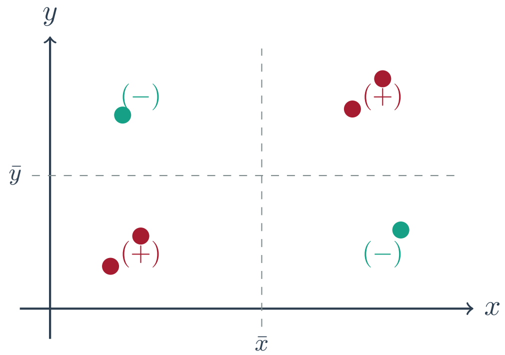
Both cases give **positive products**—they *accumulate*, not cancel!

The sum of these products tells us something:

- Positive sum $\rightarrow$ variables move together
- Negative sum $\rightarrow$ variables move opposite
- Zero sum $\rightarrow$ no systematic relationship

This sum is the heart of covariance.

The Four Quadrants



Positive covariance: Points cluster in upper-right and lower-left. **Negative**: upper-left and lower-right.

A First Attempt at the Formula

We want to measure the "average" product of deviations:

$$\textbf{WRONG: } \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Why is this wrong?

Same reason as with variance: we already used the data to estimate $\bar{x}$ and $\bar{y}$.

We've "used up" degrees of freedom, so dividing by n underestimates the true covariance.

The Covariance Formula

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

- Uses $n-1$ for same reason as variance (degrees of freedom)
- **Units**: units of x $\times$ units of y
  - Approval (%) $\times$ Population (millions) = "percent-millions"
  - This is awkward! We'll fix it later.

Why $n - 1$? The Degrees of Freedom Story

The problem: We don't know the true population means μ_x and μ_y.

We estimate them with $\bar{x}$ and $\bar{y}$—but these estimates come from the *same data* we're using to calculate covariance.

The consequence: Deviations from the sample mean are artificially small.

- The sample mean minimizes squared deviations (that's what it does!)
- So our deviations understate the true spread

The fix: Divide by $n - 1$ instead of n to correct for this bias.

This insight dates to the 1900s (William Gosset, "Student"). It's why we call it the *sample* covariance—it's designed to estimate the population covariance.

Covariance in R

```r
# Covariance between approval and population
cov(approval$approval, approval$population)
## [1] 7843521
```

The covariance is about 7.8 million... *percent-people*?

- Positive: larger states tend to have higher approval
- But is 7.8 million “big” or “small”? Hard to tell!

The units problem is why we’ll need correlation later.

Property 1: Covariance With Itself Is Variance

Compare the two formulas:

$$\textbf{Variance:}\ s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})$$

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

Key insight: If $y = x$, the covariance formula becomes the variance formula!

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

Variance is just a special case of covariance—how a variable "co-varies" with itself.

Property 2: Covariance Is Symmetric

Claim: $\text{Cov}(x, y) = \text{Cov}(y, x)$

Example: Three observations

i	x_i	y_i	$(x_i - \bar{x})(y_i - \bar{y})$	$(y_i - \bar{y})(x_i - \bar{x})$
1	2	5	$(-1)(-1) = 1$	$(-1)(-1) = 1$
2	3	6	$(0)(0) = 0$	$(0)(0) = 0$
3	4	7	$(1)(1) = 1$	$(1)(1) = 1$
Sum			**2**	**2**

$(\bar{x} = 3,\ \bar{y} = 6)$

The products are identical—multiplication is commutative! Order doesn't matter.

Property 3: Covariance Can Be Positive, Negative, or Zero

Positive covariance: Variables move together **Negative covariance**: Variables move opposite

x	y	Deviations	Product
1	2	$(-)(-)$	$+$
3	4	$(0)(0)$	0
5	6	$(+)(+)$	$+$

x	y	Deviations	Product
1	6	$(-)(+)$	$-$
3	4	$(0)(0)$	0
5	2	$(+)(-)$	$-$

$$\text{Cov}(x, y) > 0$$

$$\text{Cov}(x, y) < 0$$

When variables move together: same signs $\to$ positive products.
When variables move opposite: opposite signs $\to$ negative products.

Property 4: Zero Covariance $\neq$ No Relationship

Covariance measures *linear* relationships only.

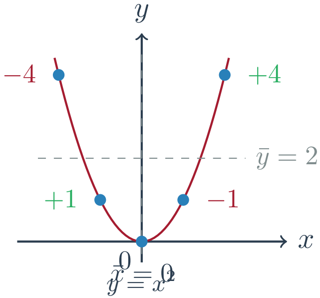
Consider $y = x^2$ with $x \in \{-2, -1, 0, 1, 2\}$:

x	$y = x^2$	$x - \bar{x}$	$y - \bar{y}$	Product
-2	4	-2	$+2$	-4
-1	1	-1	-1	$+1$
0	0	0	-2	0
1	1	$+1$	-1	-1
2	4	$+2$	$+2$	$+4$
Sum				**0**

$(\bar{x} = 0,\ \bar{y} = 2)$

$\text{Cov}(x, y) = 0$, but y is *perfectly determined* by x!

Why Zero Covariance With a Curved Relationship?



Each label $= (x_i - \bar{x})(y_i - \bar{y})$

Products sum to zero:

$$(-4) + (+1) + 0 + (-1) + (+4) = 0$$

Positive products (green) and negative products (red) cancel perfectly.

Covariance only detects **linear** relationships—not curves!

The Weighted Mean Decomposition

Now We Can Answer Our Question

Question: Why are weighted and unweighted means sometimes equal, sometimes different?

Answer: It depends on the covariance between the variable and the weights.

Let me show you why.

The Derivation: Setup

Notation:

- Weighted mean: $\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$
- Unweighted mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Mean of the weights: $\bar{w} = \frac{1}{n}\sum_{i=1}^{n} w_i$
- Sum of the weights: $W = \sum_{i=1}^{n} w_i = n\bar{w}$

Our goal: Express $\bar{x}_w$ in terms of $\bar{x}$ and covariance.

Step 1: Rewrite the Weighted Mean

Start with the weighted mean:

$$\bar{x}_w = \frac{\sum w_i x_i}{\sum w_i}$$

Since $\sum w_i = n\bar{w}$, we can write:

$$\bar{x}_w = \frac{\sum w_i x_i}{n\bar{w}}$$

Step 2: The Add-and-Subtract Trick

The classic calculus trick: add and subtract the same thing.

Add and subtract $n\bar{x}\bar{w}$ in the numerator:

$$\bar{x}_w = \frac{\sum w_i x_i - n\bar{x}\bar{w} + n\bar{x}\bar{w}}{n\bar{w}}$$

The red and blue terms cancel—we haven't changed anything!

Split into two fractions:

$$\bar{x}_w = \frac{\sum w_i x_i - n\bar{x}\bar{w}}{n\bar{w}} + \frac{n\bar{x}\bar{w}}{n\bar{w}}$$

Step 3: Simplify the Second Term

The second term simplifies immediately:

$$\frac{n\bar{x}\bar{w}}{n\bar{w}} = \bar{x}$$

So now we have:

$$\bar{x}_w = \bar{x} + \frac{\sum w_i x_i - n\bar{x}\bar{w}}{n\bar{w}}$$

The unweighted mean $\bar{x}$ is already separated out!

Step 4a: Expand the Numerator

Focus on the numerator: $\sum w_i x_i - n\bar{x}\bar{w}$

First, note that $n\bar{w} = \sum w_i$, so:

$$n\bar{x}\bar{w} = \bar{x} \cdot n\bar{w} = \bar{x}\sum w_i$$

Substitute:

$$\sum w_i x_i - n\bar{x}\bar{w} = \sum w_i x_i - \bar{x}\sum w_i$$

Factor out $\bar{x}$:

$$= \sum w_i x_i - \sum w_i \bar{x} = \sum w_i (x_i - \bar{x})$$

Step 4b: Convert to Deviation Form

We have: $\sum w_i(x_i - \bar{x})$

The trick: Write $w_i = (w_i - \bar{w}) + \bar{w}$

Substitute:

$$\sum w_i(x_i - \bar{x}) = \sum \left[(w_i - \bar{w}) + \bar{w}\right](x_i - \bar{x})$$

Expand:

$$= \sum (w_i - \bar{w})(x_i - \bar{x}) + \bar{w} \sum (x_i - \bar{x})$$

But $\sum(x_i - \bar{x}) = 0$ (deviations always sum to zero!), so:

$$\sum w_i x_i - n\bar{x}\bar{w} = \sum (w_i - \bar{w})(x_i - \bar{x})$$

This is the cross-product of deviations—the numerator of covariance!

Step 5: The Covariance Connection

We know the covariance formula:

$$\text{Cov}(x, w) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(w_i - \bar{w})$$

So:

$$\sum (x_i - \bar{x})(w_i - \bar{w}) = (n-1) \cdot \text{Cov}(x, w)$$

For large n: $(n-1)/n \approx 1$, so we can write:

$$\bar{x}_w \approx \bar{x} + \frac{\text{Cov}(x, w)}{\bar{w}}$$

The Result: The Weighted Mean Decomposition

$$\bar{x}_w = \bar{x} + \frac{\mathrm{Cov}(x, w)}{\bar{w}}$$

Weighted mean = Unweighted mean + Covariance adjustment

- If $\mathrm{Cov}(x, w) = 0$: weighted = unweighted
- If $\mathrm{Cov}(x, w) > 0$: weighted > unweighted
- If $\mathrm{Cov}(x, w) < 0$: weighted < unweighted

Back to TRANTIME

```r
# Verify the decomposition
unweighted <- mean(commuters$TRANTIME)          # 27.22
weighted <- weighted.mean(commuters$TRANTIME,
                          commuters$PERWT)      # 27.19

# The covariance term
cov_term <- cov(commuters$TRANTIME, commuters$PERWT) /
            mean(commuters$PERWT)
# cov_term ~ -0.03

# Check: 27.22 + (-0.03) = 27.19
```

The covariance between commute time and sampling weights is nearly zero.

That's why the weighted and unweighted means are almost the same!

Now Try INCTOT (Total Income)

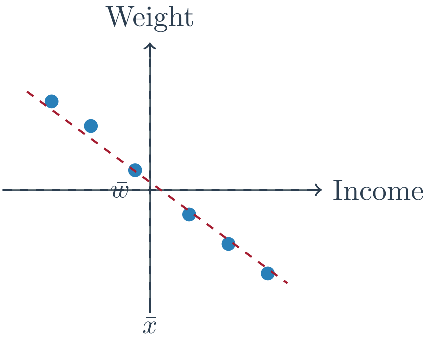
```r
# Total income from ACS
unweighted <- mean(commuters$INCTOT)          # $72,993
weighted <- weighted.mean(commuters$INCTOT,
                          commuters$PERWT)    # $69,952

# The covariance term
cov_term <- cov(commuters$INCTOT, commuters$PERWT) /
            mean(commuters$PERWT)
# cov_term ~ -3,040
```

Weighted mean is $3,000 *lower* than unweighted!

Unlike TRANTIME, here the covariance is **negative**—and it's large.

Higher Income → Lower Weights

In the ACS, **higher-income people tend to have lower sampling weights**.



Why? The Census oversamples hard-to-reach populations to ensure adequate coverage.

This Creates Negative Covariance

Income above mean $\rightarrow$ weight *below* mean $\rightarrow$ negative product

Income below mean $\rightarrow$ weight *above* mean $\rightarrow$ negative product

The products are mostly negative.

$$\boxed{\mathrm{Cov}(\mathrm{INCTOT}, \mathrm{PERWT}) \quad < \quad 0}$$

Weighting Corrects for Oversampling

	Mean Income
Unweighted (raw sample)	\$72,993
Weighted (U.S. population)	\$69,952
Difference	−\$3,041

The raw sample overrepresents higher-income people.

Weighting gives us the true U.S. population mean.

The Three Cases: A Summary

Variable	**Cov with weights**	**Weighted vs. Unweighted**
TRANTIME	≈ 0	Nearly equal
INCTOT	< 0 (negative)	Weighted *lower*
State approval	> 0 (positive)	Weighted *higher*

$$\bar{x}_w = \bar{x} + \frac{\text{Cov}(x, w)}{\bar{w}}$$

- Positive covariance $\rightarrow$ weighted mean pulled *up*
- Negative covariance $\rightarrow$ weighted mean pulled *down*
- Zero covariance $\rightarrow$ weighted $\approx$ unweighted

Back to Approval Ratings

```r
# State approval example
unweighted <- mean(approval$approval)         # 43.3
weighted <- weighted.mean(approval$approval,
                          approval$population)  # 44.5

# The covariance term
cov_term <- cov(approval$approval, approval$population) /
            mean(approval$population)
# cov_term ~ +1.2
```

Large states (CA, NY) tend to have higher approval → **positive covariance**.

That's why the weighted mean is higher than the unweighted mean!

The Takeaway

> Weighted and unweighted means differ when the covariance between the variable and the weights is non-zero.

PS1 lesson: Commute time has near-zero covariance with ACS sampling weights.

- The survey design doesn't systematically oversample long or short commuters
- So weighting doesn't change much

Political lesson: Approval has positive covariance with state population.

- Big states happen to have higher approval
- So weighting by population pulls the mean up

From Covariance to Correlation

The Problem with Covariance

Variables	Covariance
Approval & Population	7,843,521
Height (in) & Weight (lbs)	20.5

Which relationship is “stronger”?

Can’t tell! The covariances are in different units:

- Percent $\times$ people
- Inches $\times$ pounds

We need a standardized measure.

Everyday Language: “Correlation”

People say “two things are correlated” all the time.

What do they mean?

- “They move together”
- “When one goes up, the other tends to go up (or down)”

But how much?

We need a measure that:

- Has no units
- Is comparable across different variable pairs
- Has a clear interpretation

Correlation: Covariance Standardized

$$\textbf{Sample Correlation: } r_{xy} = \frac{\mathrm{Cov}(x, y)}{s_x \cdot s_y} = \frac{s_{xy}}{s_x s_y}$$

- Divide covariance by the standard deviations of both variables
- Units cancel out!
- Result is always between -1 and $+1$

Correlation is "standardized covariance."

Correlation in R

```r
# Correlation between approval and population
cor(approval$approval, approval$population)
## [1] 0.183
```

The correlation is 0.18—a weak positive relationship.

```r
# Compare to covariance
cov(approval$approval, approval$population)
## [1] 7843521
```

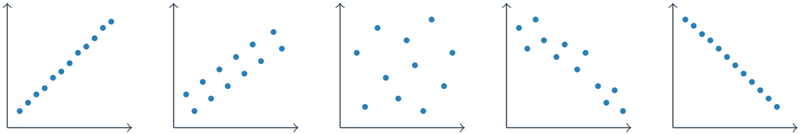
Same information, but correlation is interpretable.

Interpreting Correlation

r value	Interpretation
$+1$	Perfect positive linear relationship
$+0.7$ to $+1$	Strong positive
$+0.3$ to $+0.7$	Moderate positive
0 to $+0.3$	Weak positive
0	No linear relationship
-0.3 to 0	Weak negative
-0.7 to -0.3	Moderate negative
-1 to -0.7	Strong negative
-1	Perfect negative linear relationship

These are rules of thumb, not hard boundaries.

Visual Gallery of Correlations

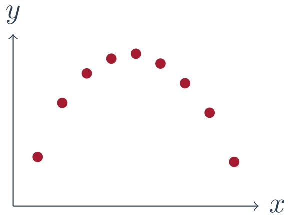


$r = 0.9$ $\quad r = 0.5$ $\quad r = 0$ $\quad r = -0.5$ $\quad r = -0.9$

As the absolute value of r increases, points cluster more tightly around a line.

Important Caveat: Linearity

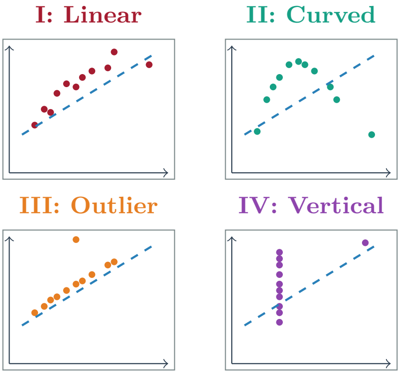
Correlation measures **linear** relationships only.



$r \approx 0$, but clear pattern!

Always plot your data! Famous example: Anscombe's quartet.

Anscombe's Quartet: Four Datasets, One Correlation



All four datasets have the same correlation: $r = 0.82$

The Lesson from Anscombe

All four datasets have: same mean ($\bar{x} = 9$, $\bar{y} = 7.5$), same variance, same correlation ($r = 0.82$), same regression line.

But the patterns are completely different:

- I: A genuine linear relationship
- II: A curved relationship (quadratic)
- III: A perfect line with one outlier
- IV: No relationship except one influential point

Summary statistics can hide important patterns.
Always visualize your data.

Correlation in Action

Research Question: Social Media and Political Knowledge

Question: Does more social media use correlate with more or less political knowledge?

This is relevant to you:

- Your generation
- Your daily experience
- A question people actually care about

Hypotheses:

- More info? Social media exposes people to news $\rightarrow$ positive r
- Distraction? Social media crowds out learning $\rightarrow$ negative r
- Misinformation? Social media spreads false beliefs $\rightarrow$???

What Would the Data Look Like?

Imagine survey data:

- x = Hours per day on social media
- y = Political knowledge score (0–100)

```
# Hypothetical analysis
cor(survey$social_media_hours, survey$political_knowledge)
## [1] -0.15
```

A weak negative correlation: people who use more social media tend to score slightly lower on political knowledge.

But what does this tell us about causation?

Correlation $\neq$ Causation

Correlation tells us variables move together.
It does NOT tell us one causes the other.

Classic example: Ice cream sales correlate with drowning deaths.

- Does ice cream cause drowning? No.
- Does drowning cause ice cream sales? No.
- Both are driven by a third variable: **summer/heat**

We'll get to causation later in the course.

What's Next?

Today: Covariance and correlation

- Describing how two variables move together
- Why weighted means differ from unweighted

Next week: Regression

- Predicting one variable from another
- Modeling relationships mathematically

Later: Causal inference

- Does X actually *cause* Y?
- When can we make causal claims?

Where We Stand: Three Core Calculations

The Building Blocks of Data Analysis

Over the past two weeks, we've learned three fundamental calculations:

Mean	**Variance**	**Covariance**
Center (location)	Spread (dispersion)	Association (co-movement)

Let's summarize the key properties of each.

Properties of the Mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

1. **Shift**: Adding a constant shifts the mean: $y_i = x_i + c \Rightarrow \bar{y} = \bar{x} + c$
2. **Scale**: Multiplying scales the mean: $y_i = a \cdot x_i \Rightarrow \bar{y} = a \cdot \bar{x}$
3. **Sum of deviations**: Deviations from the mean sum to zero:

$$\sum_{i=1}^{n}(x_i - \bar{x}) = 0$$

4. **Additivity**: The mean of a sum equals the sum of means:

$$\overline{(x + y)} = \bar{x} + \bar{y}$$

Properties of Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

1. **Always non-negative**: $s^2 \geq 0$ (zero only if all values identical)
2. **Shift invariant**: Adding a constant doesn't change variance: $y_i = x_i + c \Rightarrow s_y^2 = s_x^2$
3. **Scale squared**: Multiplying by a multiplies variance by a^2:

$$y_i = a \cdot x_i \Rightarrow s_y^2 = a^2 \cdot s_x^2$$

4. **Standard deviation**: $s = \sqrt{s^2}$ returns to original units

Properties of Covariance

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

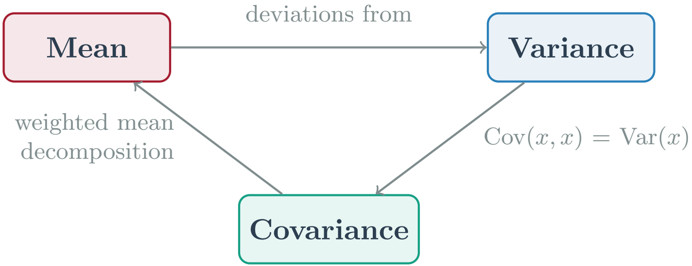
1. **Symmetry**: Order doesn't matter: $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$
2. **Self-covariance is variance**: $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$
3. **Shift invariant**: Adding constants doesn't change covariance:

$$\mathrm{Cov}(x + a, y + b) = \mathrm{Cov}(x, y)$$

4. **Scale**: Multiplying by constants scales covariance:

$$\mathrm{Cov}(ax, by) = ab \cdot \mathrm{Cov}(x, y)$$

The Connections Between Them



- Variance uses deviations from the mean
- Covariance generalizes variance to two variables
- Covariance explains when weighted $\neq$ unweighted

Recap

Today's Key Concepts

1. **Covariance**: Do two variables move together?

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

2. **Weighted mean decomposition**:

$$\bar{x}_w = \bar{x} + \frac{\text{Cov}(x, w)}{\bar{w}}$$

3. **Correlation**: Standardized covariance (-1 to $+1$)

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

For Problem Set 2

You'll:

- Calculate covariance and correlation
- Interpret what they mean substantively
- Connect this to the regression we'll learn next week

R functions to know:

- `cov(x, y)` — covariance
- `cor(x, y)` — correlation
- `weighted.mean(x, w)` — weighted mean

When weighted $\neq$ unweighted,
look for covariance between
the variable and the weights.

Questions?