The CEF, Sampling, and Estimation

Gov 2001: Quantitative Social Science Methods I

Scott Cunningham

Harvard University

Spring 2026

Today's Reading

Required

- **Aronow & Miller**, §2.2.3–2.2.4: CEF, LIE, best predictor (pp. 72–88)
- **Aronow & Miller**, Ch. 3: Learning from Random Samples (pp. 115–138)
- **Blackwell**, Ch. 1: What is regression really doing?

Two big moves today: First, we learn the CEF—the function that regression is trying to estimate. Then we ask: how do we learn about populations from *samples*?

CEF $\longrightarrow$ sampling $\longrightarrow$ plug-in estimation. This is the logic behind everything from Wooldridge's "sample regression function" to OLS itself.

From Conditional Distributions to Conditional Means

A short history of the most important function in statistics

Last week we learned conditional distributions: $f_{Y|X}(y|x)$ captures everything about how Y relates to X.

But a full distribution is a lot of information.

In 1805, **Adrien-Marie Legendre** published the method of least squares. His question was practical: given noisy astronomical observations, how do you find the best-fitting curve?

His answer—minimize squared prediction errors—implicitly targets the **conditional mean**. The CEF was hiding inside regression for 200 years before anyone named it.

Legendre didn't think in terms of conditional distributions. The formal connection came much later, through Kolmogorov (1933) and the modern probability framework.

The Practical Question

You're an analyst at a campaign. Your boss asks:

"Among voters with a college degree, what's the average level of support for our candidate?"

What your boss wants is: $\mathbb{E}[\text{Support} \mid \text{Education} = \text{College}]$

She doesn't want:

- The full distribution of support among college voters
- Just the overall average support
- A complicated model

She wants a single number that summarizes support, conditional on education.

The Conditional Expectation Function

Definition

The **Conditional Expectation Function** (CEF) is:

$$G_Y(x) = \mathbb{E}[Y|X = x]$$

What is this?

- For each value of x, compute the expected value of Y among units with $X = x$
- The result is a *function* of x
- It summarizes the conditional distribution with a single number

Other names: Conditional mean, regression function

Blackwell calls this "the thing regression is trying to estimate."

Computing the CEF

The formulas you need

For continuous Y:

$$\mathbb{E}[Y|X=x] = \int_{-\infty}^{\infty} y \cdot f_{Y|X}(y \mid x)\, dy$$

For discrete Y:

$$\mathbb{E}[Y|X=x] = \sum_{y} y \cdot \Pr(Y = y \mid X = x)$$

Key point: The CEF is a *function of x*—plug in different values of x and you get different numbers. It's not a single number.

We learned conditional distributions last week. The CEF just takes their expected value.

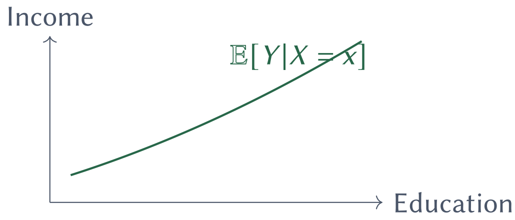
Example: Wages and Education

Setup: Y = annual income, X = years of education

The CEF $G_Y(x) = \mathbb{E}[\text{Income}|\text{Education} = x]$ answers:

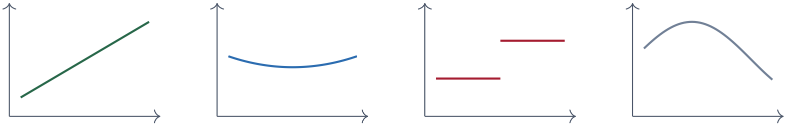
- What's the average income among people with 12 years of education?
- What's the average income among people with 16 years?
- What's the average income among people with 20 years?

The CEF traces out how average income changes with education.



The CEF Can Be Any Shape

Nothing requires the CEF to be linear.



Linear Quadratic Step Nonmonotonic

Regression typically assumes linearity: $\mathbb{E}[Y|X = x] = \alpha + \beta x$
This is a *modeling assumption*, not a fact about the world.

When we get to OLS, we'll see it as approximating the true CEF with a line.

Why the CEF Matters: Best Prediction

Claim: The CEF is the *best predictor* of Y given X.

What do we mean by "best"?
Suppose you must predict Y using only X. You choose some function $g(X)$.

Define the **Mean Squared Error** of your prediction:

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

Theorem: CEF is the MSE-Optimal Predictor

Among *all* functions $g(X)$, the CEF minimizes MSE:

$$\mathbb{E}[Y|X] = \arg\min_{g(X)} \mathbb{E}\left[(Y - g(X))^2\right]$$

Intuition: Why the CEF is Best

Think about what you're doing when you predict Y from X:

1. You observe $X = x$
2. You know the distribution of Y given $X = x$
3. You need to pick a single number as your guess

We already proved (Week 3): The best constant predictor of a random variable is its expected value.

Applying that here: Once we condition on $X = x$, the best prediction of Y is $\mathbb{E}[Y|X = x]$.

The CEF is just "pick the mean" applied separately for each $X = x$.

The CEF Residual

Define the CEF residual:

$$\varepsilon = Y - \mathbb{E}[Y|X]$$

This is what's "left over" after the CEF prediction.

Key Property of CEF Residuals

$$\mathbb{E}[\varepsilon|X] = 0$$

Why?

$$\begin{aligned} \mathbb{E}[\varepsilon|X] &= \mathbb{E}[Y - \mathbb{E}[Y|X] \mid X] \\ &= \mathbb{E}[Y|X] - \mathbb{E}[Y|X] = 0 \end{aligned}$$

The residual has mean zero *at every value of X*, not just overall.

The Foundational Property: Orthogonality

CEF Residual Orthogonality

$$\text{Cov}(\varepsilon, g(X)) = 0 \quad \text{for any function } g$$

In words: The CEF residual is uncorrelated with *any* function of X.

Why this matters:

- There is no remaining systematic relationship with X
- No transformation of X could improve the prediction
- This is the property regression tries to achieve

Regression residuals will satisfy a weaker version: $\text{Cov}(u, X) = 0$ (just linear).

The CEF Decomposition

We can always write:

$$Y = \mathbb{E}[Y|X] + \varepsilon$$

where $\mathbb{E}[\varepsilon|X] = 0$.

This is a **decomposition** of Y into:

- **Systematic part**: $\mathbb{E}[Y|X]$ — what X predicts
- **Idiosyncratic part**: ε — unpredictable from X

Regression does the same thing, but with a linear approximation:

$$Y = \alpha + \beta X + u$$

We'll make this connection precise when we cover OLS.

The Law of Iterated Expectations (LIE)

Law of Iterated Expectations

$$\mathbb{E}[Y] = \mathbb{E}\left[\mathbb{E}[Y|X]\right]$$

In words: The overall mean of Y equals the average of the conditional means, weighted by the distribution of X.

Discrete case:

$$\mathbb{E}[Y] = \sum_x \mathbb{E}[Y|X = x] \cdot \Pr(X = x)$$

Also called the "law of total expectation" or "tower property."

LIE Example: Average Wages

Setup: Two groups—college grads and non-college.

Group	Share	Avg Wage
Non-College	0.60	$45,000
College	0.40	$75,000

What's the overall average wage?

Using LIE:

$$\begin{aligned}\mathbb{E}[\text{Wage}] &= \mathbb{E}[\text{Wage}|\text{No College}] \cdot \Pr(\text{No College}) \\ &\quad + \mathbb{E}[\text{Wage}|\text{College}] \cdot \Pr(\text{College}) \\ &= 45{,}000 \times 0.60 + 75{,}000 \times 0.40 \\ &= 27{,}000 + 30{,}000 = \$57{,}000\end{aligned}$$

LIE is Everywhere in Statistics

You'll use this constantly:

- Proving unbiasedness of estimators
- Deriving variance decompositions
- Understanding omitted variable bias
- Causal inference (potential outcomes, weighting)

Example preview (OVB derivation):
"What's the expected value of the short regression coefficient?"
"First condition on X, compute the expectation, then average over X."

Mastering LIE is essential for the rest of this course.

LIE with Extra Conditioning

The general version you'll need for proofs

Standard LIE (what we just saw):

$$\mathbb{E}[Y] = \mathbb{E}\big[\mathbb{E}[Y|X]\big]$$

With extra conditioning on Z:

$$\mathbb{E}[Y|Z] = \mathbb{E}\big[\mathbb{E}[Y|X,Z] \,\big|\, Z\big]$$

"Average first over X (holding Z fixed), then you have a function of Z only."

Conditioning on functions: If g is any function of X, then

$$\mathbb{E}[Y|g(X),X] = \mathbb{E}[Y|X]$$

Adding $g(X)$ provides no new information beyond X itself.

Example: If you know income (X), also knowing tax bracket ($g(X)$) doesn't help predict consumption.

The Variance Decomposition

Another use of the CEF: Decomposing variance.

Law of Total Variance

$$\text{Var}(Y) = \mathbb{E}[\text{Var}(Y|X)] + \text{Var}(\mathbb{E}[Y|X])$$

In words:

- Total variance = Within-group variance + Between-group variance
- $\mathbb{E}[\text{Var}(Y|X)]$ = Average variance of Y within each X group
- $\text{Var}(\mathbb{E}[Y|X])$ = Variance of the group means

This is the foundation of R-squared in regression.

Example: Wage Variance

Setup: Same as before, but now with within-group variance.

Group	Share	Mean Wage	SD of Wage
Non-College	0.60	\$45,000	\$15,000
College	0.40	\$75,000	\$25,000

Within-group variance: $\mathbb{E}[\text{Var}(Y|X)]$

$$= 0.60 \times (15{,}000)^2 + 0.40 \times (25{,}000)^2 = 385{,}000{,}000$$

Between-group variance: $\text{Var}(\mathbb{E}[Y|X])$

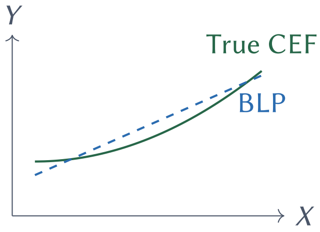
$$= 0.60 \times (45{,}000 - 57{,}000)^2 + 0.40 \times (75{,}000 - 57{,}000)^2 = 216{,}000{,}000$$

Total variance: $385M + 216M = 601{,}000{,}000$

Blackwell's Take: CEF vs. Linear Regression

From Blackwell (Ch. 1):

Linear regression finds the best *linear* approximation to the CEF, whatever shape the CEF has.



Key insight: Regression doesn't *assume* the CEF is linear. It finds the line that gets closest to the true CEF.

When the CEF *is* linear (as in the bivariate normal), the BLP and the CEF coincide.

Applications in Political Science

The CEF is everywhere in political science research:

- $\mathbb{E}[\text{Vote Share}|\text{Incumbent}]$: Average vote share for incumbents vs. challengers
- $\mathbb{E}[\text{Turnout}|\text{Age}]$: How turnout varies with age
- $\mathbb{E}[\text{Approval}|\text{Economy}]$: Presidential approval as a function of economic conditions
- $\mathbb{E}[\text{Policy Position}|\text{Party}]$: Average policy positions by party

Regression estimates these relationships from data.

From Population to Sample: The Bridge

So far today: We defined the CEF as a *population* quantity.

- $\mathbb{E}[Y|X]$ is a function of the population distribution
- The LIE, orthogonality, and variance decomposition are all population results
- Regression *targets* the CEF

The problem: We don't observe the population. We have a *sample*.

How do we estimate population quantities from sample data?

This is the fundamental question of statistics—and the topic for the rest of today.

The Fundamental Problem of Statistics

What we want: Population parameters

- Population mean: $\mu = \mathbb{E}[Y]$
- Population variance: $\sigma^2 = \mathrm{Var}(Y)$
- Conditional expectation: $\mathbb{E}[Y|X]$

What we have: A sample of n observations

- $Y_1, Y_2, \ldots, Y_n$ drawn from the population

The question: Can we use the sample to learn about the population?

Yes—under the right conditions. That's what this week is about.

A Brief History of Estimation

- **1800s**: Gauss and Laplace develop least squares for *astronomy*—parameters are fixed constants to recover (planet orbits, not social phenomena)
- **1900s**: Karl Pearson fits distributions to data, implicitly assuming populations exist with fixed parameters
- **1922**: R.A. Fisher formalizes the framework we use today—population has fixed parameters, samples estimate them
- **1923**: Neyman proposes an alternative: inference from *randomization*, not from assuming a superpopulation
- **1950**: Wald frames statistics as a *decision problem*—bias, variance, and risk become the criteria

Key shift: From "how do I fit this curve?" to "what can I learn about the population from a sample?"

Not Everyone Agreed

The "population first" view won—but alternatives existed

Bayesians (Laplace 1774, Jeffreys 1939, Savage 1954):

- Parameters have *distributions* reflecting uncertainty, not fixed values

Design-based (Neyman 1923, Freedman 2008):

- Inference from randomization, not from assuming a superpopulation

Classical econometrics (Goldberger 1991, Amemiya 1985):

- Coefficients are fixed unknowns. Randomness is only in ε, never in β.

Today's framework: A&M call i.i.d. sampling a "codification of uncertainty about generalizability." The population is a **useful fiction**—not a discovered truth.

This Debate Changed How We Do Empirical Work

The old way: Run a regression, report $\hat{\beta}$, call it "the effect."

The problem: $\hat{\beta}$ *of what population*? Identified *how*?

The modern way: Define your *estimand* first (what are you trying to learn?), *then* show your estimator hits it under stated assumptions.

Today we formalize this logic: population quantity $\longrightarrow$ sample analog $\longrightarrow$ evaluate the estimator.

Wooldridge's "population regression vs. sample regression" is this same idea applied to OLS.

Motivating Example: Does Social Pressure Increase Turnout?

Gerber, Green, and Larimer (APSR, 2008)

Experiment: Mail voters a letter showing their neighbors' voting history.

Data: 344,084 registered voters in Michigan, randomly assigned to treatment groups.

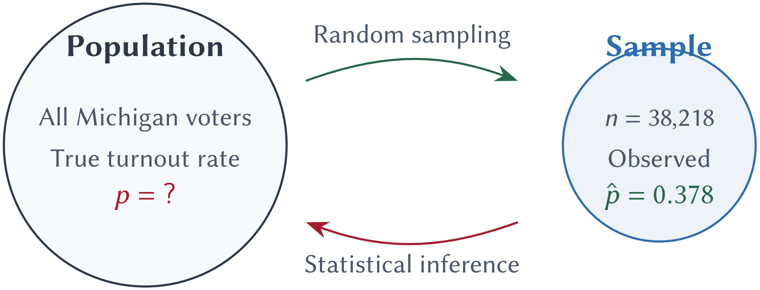
Results:

- "Neighbors" treatment: $\bar{Y}_{\text{neighbors}} = 0.378$
- "Civic Duty" control: $\bar{Y}_{\text{civic}} = 0.315$
- Difference: $0.378 - 0.315 = 0.063$

The question: Is this difference *real*? Or could it just be sampling noise?

To answer this, we need to understand how estimators behave across repeated samples.

From Population to Sample



Estimand: p
(the truth)

Estimator: $\hat{p} = \bar{X}_n$
(the formula)

Estimate: 0.378
(the number)

From Finite Population to Random Sample

Aronow & Miller, §3.1

Start concrete: A finite population of N units with values $x_1, \ldots, x_N$.

Population mean: $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ Population variance: $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$

Random sampling: Draw n units at random. Each draw has the *same* distribution (identically distributed) and draws don't affect each other (independent).

I.I.D.: $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F$

1. **Independent**: $X_i \perp\!\!\!\perp X_j$ for all $i \neq j$
2. **Identically distributed**: Each X_i has the same CDF F

A&M: The i.i.d. assumption is a "codification of uncertainty about generalizability." It's an approximation—it fails for clustered data, time series, or convenience samples.

Formalizing the Framework

Precise definitions for the visual we just saw

Estimand $\theta = T(F)$:

- A function of the population distribution—the *target*
- Examples: $\mu = \mathbb{E}[X]$, $\sigma^2 = \text{Var}(X)$, treatment effect

Estimator $\hat{\theta}_n = h(X_1, \ldots, X_n)$:

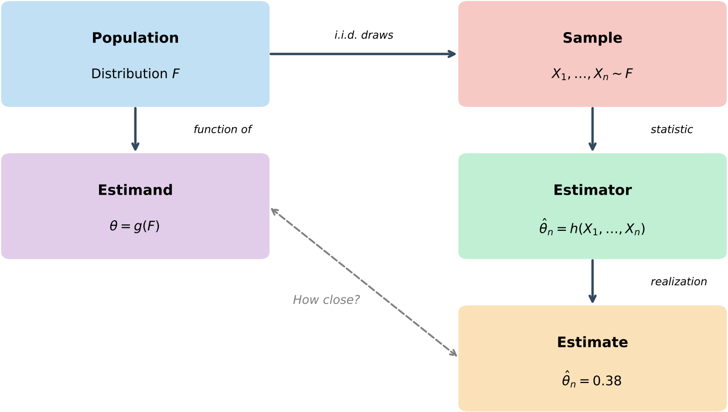
- A function of the sample—a **random variable** (it has a distribution!)

Estimate:

- A realized value of the estimator—a *number*, not a random variable

Common mistake: "My estimator was 0.38." No—your *estimate* was 0.38. The estimator is the formula.

The Estimation Framework



Many Estimators, One Estimand

Which one should we use?

Estimand: $\mu = \mathbb{E}[X]$ (the population mean)

Possible estimators:

1. $\hat{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ (sample mean)
2. $\hat{\theta}_n = X_1$ (just the first observation)
3. $\hat{\theta}_n = \max(X_1, \ldots, X_n)$ (the maximum)
4. $\hat{\theta}_n = 3$ (always guess 3)

All of these are functions of the sample. All are "estimators."

But they are not equally good. How do we choose?

We need criteria for evaluating estimators. That's what finite sample properties give us.

The Sample Mean as an Estimator

Natural idea: Estimate μ with the sample mean:

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

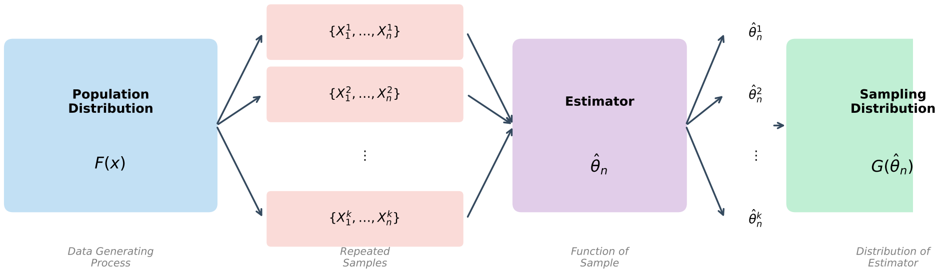
Key insight: $\bar{X}_n$ is itself a *random variable*.

- Different samples give different values of $\bar{X}_n$
- $\bar{X}_n$ has its own distribution—the **sampling distribution**
- We want to understand this distribution

In the Gerber/Green experiment: $\bar{X}_n = 0.378$ is one draw from the sampling distribution of the sample mean.

Statistics is about understanding how estimators behave across repeated samples.

The Three Distributions



The Three Distributions: What Are They?

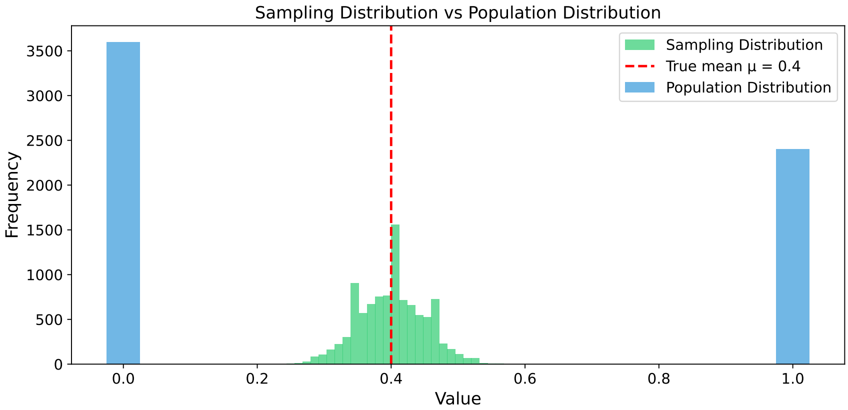
1. Population Distribution $F(x)$: The true DGP (unknown). In the voter experiment: Bernoulli(p).

2. Empirical Distribution: The observed sample $X_1, \ldots, X_n$—what we actually have. A series of 1s and 0s.

3. Sampling Distribution $G(\hat{\theta}_n)$: Distribution of the estimator over repeated samples. The 0.378 sample mean is *one draw* from this distribution.

Trick question: "The sampling distribution is the distribution of θ." True or false? False—it's the distribution of $\hat{\theta}_n$, the estimator!

Sampling Distribution vs. Population Distribution



Population is binary (0s and 1s). But the sampling distribution of $\bar{X}_n$ is bell-shaped and concentrated around the true mean.

Where Do Estimators Come From?

Two main approaches:

1. **Parametric Modeling**:
   - Assume F belongs to a known family (e.g., Normal, Poisson)
   - Use **maximum likelihood** to estimate parameters
   - Downside: inferences are model-dependent
   - (We'll cover MLE in a few weeks)

2. **Nonparametric / Plug-in**:
   - Make minimal assumptions on F
   - Replace F with the empirical distribution $\hat{F}_n$
   - More robust, fewer assumptions

The Plug-in Principle

A&M, §3.2.6: "Just pretend the data IS the population"

Idea: Replace the unknown F with the empirical distribution $\hat{F}_n$.

Empirical CDF:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$$

Plug-in estimator: If $\theta = T(F)$, then $\hat{\theta} = T(\hat{F}_n)$.

Whatever operation you'd do on the population, do it on the sample.

$$\mathbb{E}[h(X)] \quad \rightsquigarrow \quad \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

Plug-in Estimators: Examples

Sample Mean:

$$\mu = \mathbb{E}[X] \quad \rightsquigarrow \quad \hat{\mu} = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

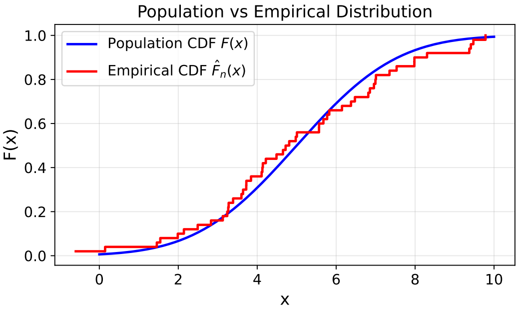
Sample Variance:

$$\sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] \quad \rightsquigarrow \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Sample Covariance:

$$\sigma_{xy} = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \quad \rightsquigarrow \quad \hat{\sigma}_{xy} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

Plug-in Principle: Visualization



Plug-in Principle: Replace F with $\hat{F}_n$

Quantity	Population	Plug-in Estimator
Mean	$\mu = E[X]$	$\hat{\mu} = \bar{X}_n$
Variance	$\sigma^2 = E[(X-\mu)^2]$	$\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$
CDF at x	$F(x) = P(X \leq x)$	$\hat{F}_n(x) = \frac{1}{n}\sum I(X_i \leq x)$

You Already Know These Estimators

The plug-in principle gives a name to what you've been doing

Everything from Weeks 2–4 was about *population* quantities:

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

The plug-in principle says: to *estimate* these, replace population expectations with sample averages.

Quantity	**Population**	**Sample (Plug-in)**
Mean	$\mu = \mathbb{E}[X]$	$\bar{X}_n$
Variance	$\sigma^2 = \text{Var}(X)$	$\frac{1}{n}\sum(X_i - \bar{X})^2$
Covariance	$\text{Cov}(X, Y)$	$\frac{1}{n}\sum(X_i - \bar{X})(Y_i - \bar{Y})$

This isn't new math. It's a framework for understanding what you've been doing.

How Do We Know If an Estimator Is Good?

Two types of properties:

- **Finite sample**: Properties for a fixed sample size n
  - Bias
  - Variance
  - Mean Squared Error (MSE)
- **Large sample (asymptotic)**: Properties as $n \to \infty$
  - Consistency (Law of Large Numbers)
  - Asymptotic normality (Central Limit Theorem)
  - *(We'll cover these later!)*

Today: Finite sample properties. These hold for *any* n.

Bias

Definition

$$\text{Bias}(\hat{\theta}_n) = \mathbb{E}[\hat{\theta}_n] - \theta$$

Unbiased: $\text{Bias}(\hat{\theta}_n) = 0$, i.e., $\mathbb{E}[\hat{\theta}_n] = \theta$.

Is the sample mean unbiased?

$$\mathbb{E}[\bar{X}_n] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n\mu = \mu$$

Yes! On average (over repeated samples), $\bar{X}_n$ hits the true value.

Unbiasedness is preserved under linear transformations. What about a weighted average?

Unbiased $\neq$ Consistent

A&M Theorem 3.2.16: Neither implies the other

Consider estimating $\mu = \mathbb{E}[X]$ with just the first observation: $\hat{\mu} = X_1$

- **Unbiased?** Yes: $\mathbb{E}[X_1] = \mu \quad \checkmark$
- **Consistent?** No: X_1 doesn't improve with more data!

Compare with the sample mean: $\hat{\mu} = \bar{X}_n$

- Unbiased *and* consistent—it uses all the data

Unbiasedness says "right on average." Consistency says "gets closer as n grows." Different virtues.

We'll formalize consistency when we cover the Law of Large Numbers.

Variance of the Estimator

Sampling variance: How spread out is $\hat{\theta}_n$ around its mean?

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

For the sample mean:

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \text{Var}(X_i) = \frac{\sigma^2}{n}$$

Key insight: Variance shrinks as n increases—more data means more precision.

Standard Error: Precision in Practice

Standard Error: The standard deviation of the estimator.

$$\text{SE}(\hat{\theta}_n) = \sqrt{\text{Var}(\hat{\theta}_n)}$$

For the sample mean: $\text{SE}(\bar{X}_n) = \sigma/\sqrt{n}$

How fast does precision improve?

- $n = 100$: $\text{SE} = \sigma/10$
- $n = 10{,}000$: $\text{SE} = \sigma/100$

To cut SE in half, you need 4× the sample size. (Why? Because $\sqrt{4n} = 2\sqrt{n}$.)

Standard Error in Practice

Back to Gerber, Green, and Larimer

In the "Neighbors" treatment group ($n = 38{,}218$, Bernoulli):

$$\mathrm{SE}(\hat{p}) = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.378 \times 0.622}{38{,}218}} \approx 0.0025$$

The sample mean is typically within 0.25 percentage points of the true turnout rate.

The treatment effect (0.063) is about 25× the SE.
That's not sampling noise—that's a real effect.

We'll formalize this "many times the SE" reasoning when we get to hypothesis testing.

Mean Squared Error (MSE)

Definition

$$\text{MSE}(\hat{\theta}_n) = \mathbb{E}[(\hat{\theta}_n - \theta)^2]$$

MSE measures how far the estimator is from the true parameter, on average.

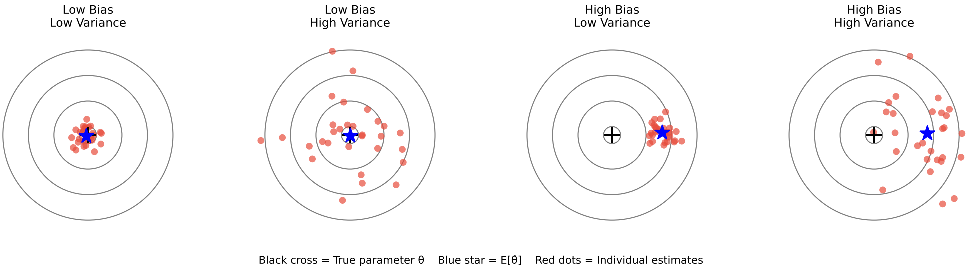
Key decomposition:

$$\text{MSE} = \text{Bias}^2 + \text{Var}$$

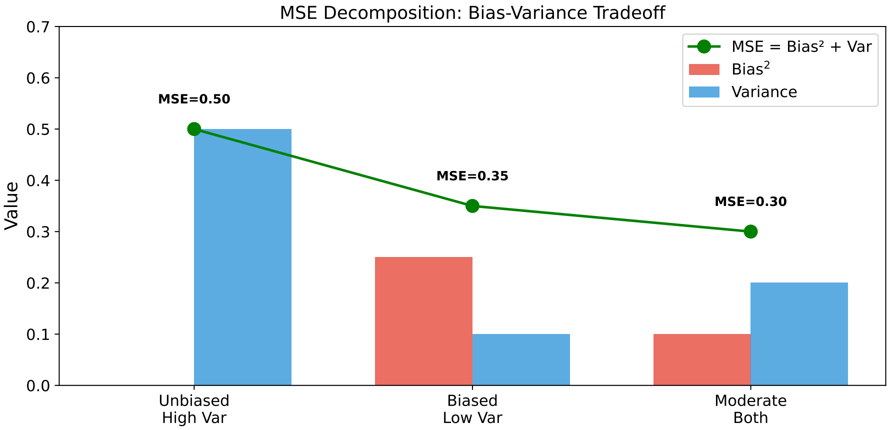
Implications:

- For unbiased estimators: MSE = Var
- Sometimes we accept *some* bias for much lower variance $\Longrightarrow$ lower MSE
- This is the **bias-variance tradeoff**

The Bias-Variance Tradeoff



MSE Decomposition: An Example



The unbiased estimator has the highest MSE! Bias isn't everything.

Worked Example: Comparing Two Estimators

The kind of problem you'll see on exams

Setup: X_1, X_2, X_3 i.i.d. from F with $\mathbb{E}[X] = \mu$, $\text{Var}(X) = \sigma^2 = 4$.

Estimator A: $\hat{\mu}_A = \bar{X}_3 = \frac{X_1+X_2+X_3}{3}$

Estimator B: $\hat{\mu}_B = \frac{1}{2}X_1 + \frac{1}{4}X_2 + \frac{1}{4}X_3$

Bias: Both have $\mathbb{E}[\hat{\mu}] = \mu$ (check the weights sum to 1). Both unbiased.

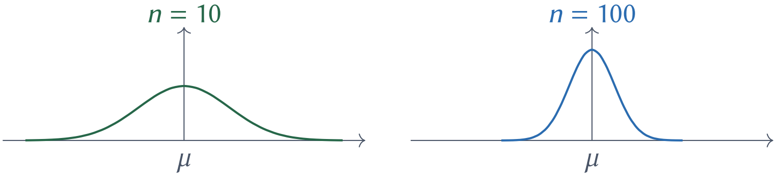
Variance:

$$\text{Var}(\hat{\mu}_A) = \frac{\sigma^2}{3} = \frac{4}{3} \approx 1.33 \qquad \text{Var}(\hat{\mu}_B) = \left(\frac{1}{4} + \frac{1}{16} + \frac{1}{16}\right)\sigma^2 = \frac{3}{8} \cdot 4 = 1.5$$

MSE: Since both unbiased, MSE = Var. **Estimator A wins** (1.33 < 1.5).

Equal weights minimize variance among all unbiased linear estimators.

Larger Samples ⇒ Lower Variance ⇒ Lower MSE



Both centered at μ (unbiased). Larger $n \Rightarrow \text{Var}(\bar{X}_n) = \sigma^2/n$ shrinks ⇒ **sampling distribution collapses toward μ.**

Key Takeaways

1. **The CEF** $\mathbb{E}[Y|X = x]$ is the best predictor of Y given X; regression approximates it
2. **LIE**: $\mathbb{E}[Y] = \mathbb{E}[\mathbb{E}[Y|X]]$—average the conditional averages
3. **Population first**: Define the estimand $\theta = T(F)$ *before* choosing an estimator
4. **Plug-in principle**: Replace F with $\hat{F}_n$—whatever you'd do on the population, do on the sample
5. **Bias**: $\mathbb{E}[\hat{\theta}] - \theta$. Unbiased $\neq$ consistent.
6. **MSE** = $\text{Bias}^2 + \text{Var}$. Sometimes a little bias is worth a lot less variance.

Looking Ahead

Next time: Asymptotics—the Law of Large Numbers and Central Limit Theorem

- What happens as $n \to \infty$? (Consistency, convergence)
- What *shape* does the sampling distribution take? (CLT: it's Normal!)
- This is the foundation for confidence intervals and hypothesis tests

Coming soon: Maximum Likelihood Estimation—a principled way to construct estimators when you have a model.

Midterm will focus on material before asymptotics—everything through today.