Estimation Theory

Properties of Estimators, Sufficiency, UMVUE, and GMM

Three Roads to the Same Estimator

Big picture: least squares, maximum likelihood, and Bayesian estimation all answer a version of the same question — *which parameter value fits the data best?* They just define "best" differently.

- **Least squares (LS):** pick θ that minimizes the total squared prediction error
- **Maximum likelihood (MLE):** pick θ that makes the observed data most probable
- **Bayesian (MAP):** pick θ that is most probable *given the data and our prior beliefs*

Key Insight

Under the right assumptions, these three different ideas lead to the **exact same estimator**. The next two slides show why.

From Maximum Likelihood to Least Squares

Suppose $Y = f(X;\theta) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$.

The likelihood of the data is a product of normal densities:

$$\text{MLE} = \arg\max_{\theta} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - f(x_i;\theta))^2}{2\sigma^2}}$$

Taking logs and dropping the parts that don't involve θ...

$$\text{MLE} = \arg\min_{\theta} \sum_{i=1}^{n} (y_i - f(x_i;\theta))^2 = \textbf{Least Squares}$$

Takeaway

If the errors are normal, least squares *is* maximum likelihood – no coincidence that both are so popular.

From Bayesian MAP to Regularized Least Squares

With a prior belief $p(\theta)$ about the parameter, the **MAP estimator** maximizes the posterior distribution:

$$\text{MAP} = \arg\max_{\theta} p(\theta \mid y) = \arg\max_{\theta} \; p(y \mid \theta) \cdot p(\theta)$$

The *shape* of the prior determines which familiar method you end up with:

Gaussian prior

$$p(\theta) \propto e^{-\lambda \|\theta\|^2}$$

$$\text{MAP} = \textbf{Ridge Regression}$$

Laplace prior

$$p(\theta) \propto e^{-\lambda \|\theta\|_1}$$

$$\text{MAP} = \textbf{LASSO}$$

Takeaway

“Regularization” in machine learning is a Bayesian prior in disguise.

Properties of Estimators

What Makes an Estimator "Unbiased"?

Intuition: imagine repeating your study many times, each time computing $\hat{\theta}$ from a fresh sample.

- If the *average* of all those estimates equals the true θ, the estimator is **unbiased**.
- If it systematically overshoots or undershoots, it's **biased** — like a bathroom scale that always reads 2 kg too heavy.

Formal Definition

$\hat{\theta}$ is **unbiased** for θ if

$$\mathrm{E}[\hat{\theta}] = \theta \quad \text{for all } \theta \in \Theta$$

Its **bias** is $\mathrm{Bias}(\hat{\theta}) = \mathrm{E}[\hat{\theta}] - \theta$.

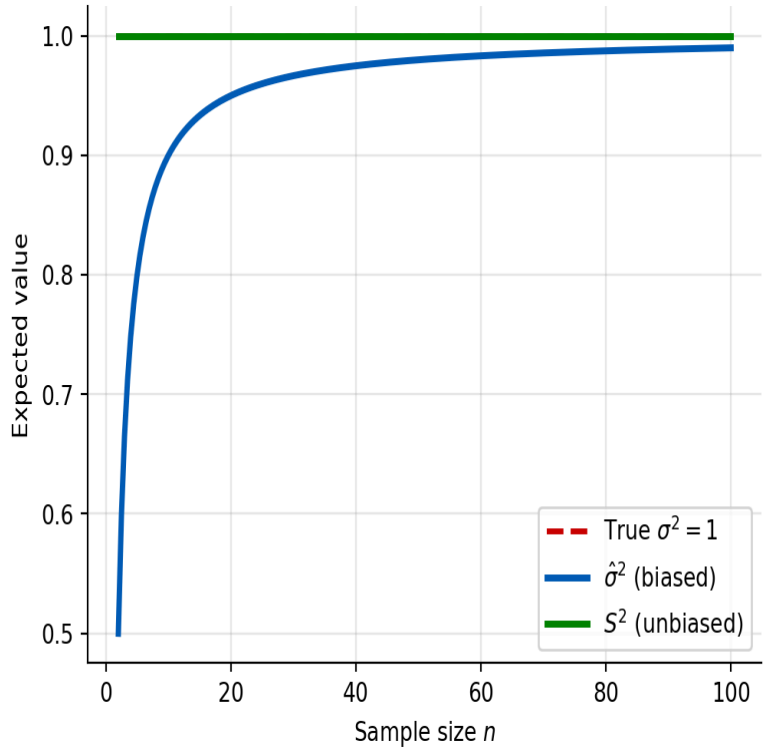
Unbiased or Biased? Some Familiar Examples

Unbiased

- $\bar{X}$ for the mean μ
- $S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ for σ^2

Biased

- $\hat{\sigma}^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$
- Most MLEs, in finite samples



Behind the Scenes: This Is Live Python

That last plot wasn't a static image — it's matplotlib, executed when the deck renders:

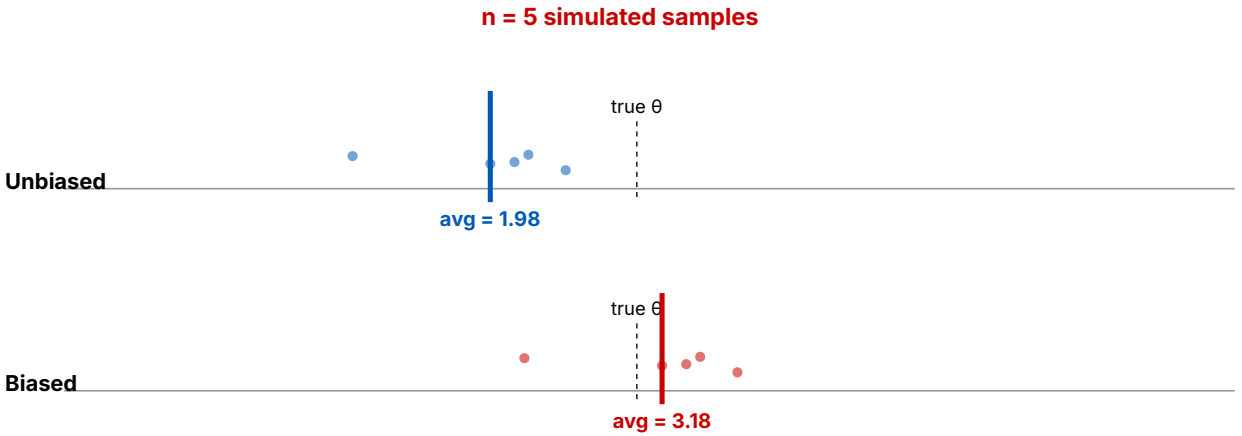
```
n = np.linspace(2, 50, 200)
fig, ax = plt.subplots(figsize=(5.4, 4.2))
ax.plot(n, np.ones_like(n), color=MYRED, linestyle='--', label=r'True $\sigma^2=1$')
ax.plot(n, (n - 1) / n, color=MYBLUE, linewidth=2.6, label=r'$\hat\sigma^2$ (biased)')
ax.plot(n, np.ones_like(n), color=MYGREEN, linewidth=2.6, label=r'$S^2$ (unbiased)')
```

Why this matters

Change the formula, re-render, and the figure updates itself — same idea works for R (`{r}` chunks) or a genuinely interactive widget instead of a static figure.

Unbiasedness, Visualized

Each dot is the estimate from one (simulated) repeated sample. Watch where the dots — and their running average — land relative to the true θ as you add more samples.



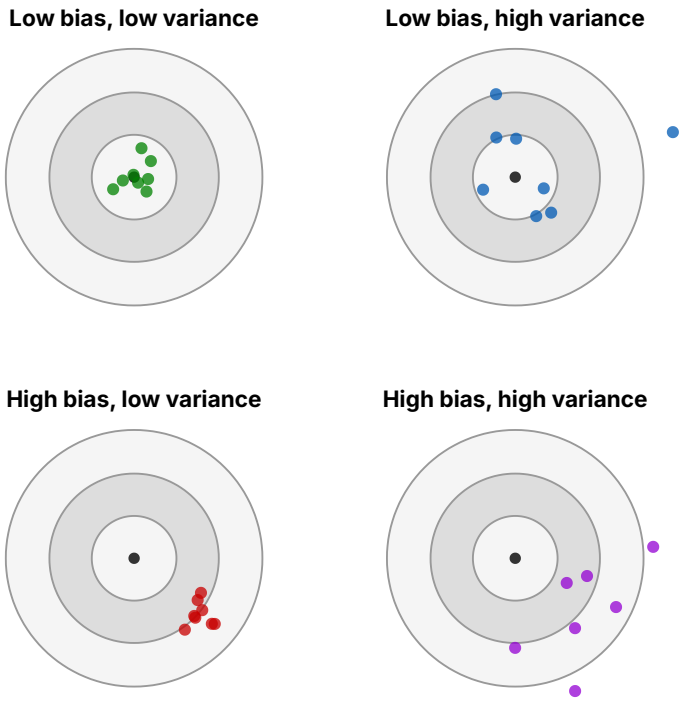
Bias vs. Variance: Two Ways to Be Wrong

Think of an estimator as arrows thrown at a target, where the bullseye is the true θ:

- **High bias:** the arrows cluster tightly, but off to one side — consistently wrong in the same way
- **High variance:** the arrows scatter widely around the bullseye — right *on average*, but unpredictable in any one sample
- An ideal estimator has *both* low bias and low variance — but in practice we often have to trade one off against the other

Bias and Variance, Visualized

Drag the slider to fire more shots at each target and see the pattern emerge.



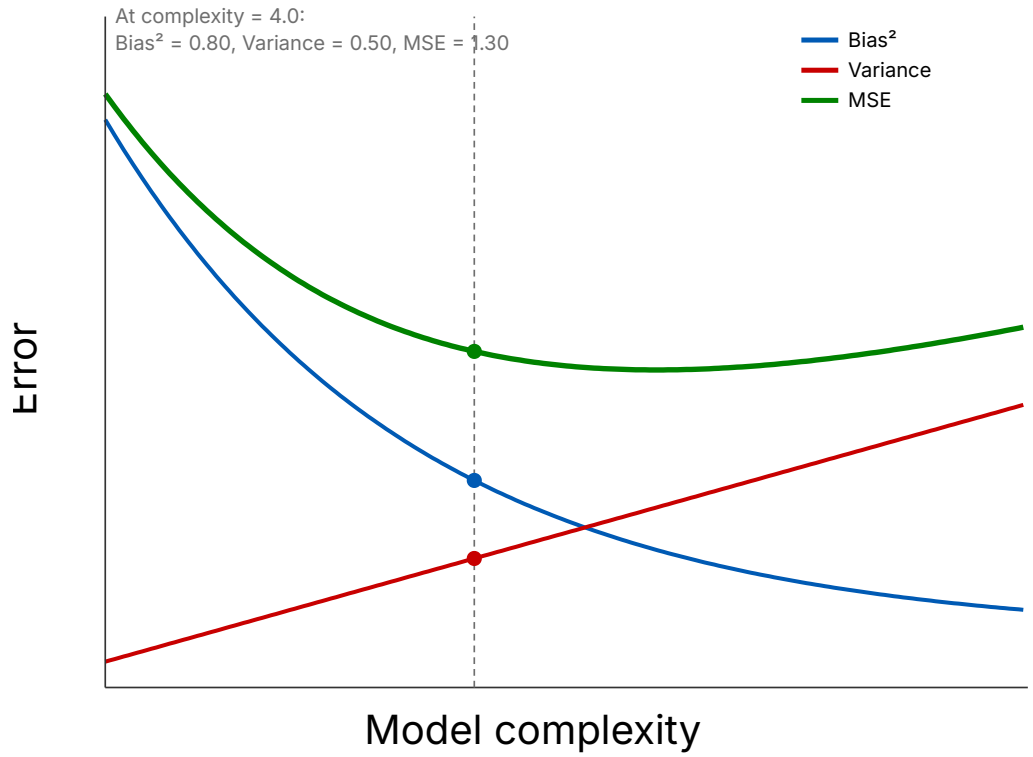
Mean Squared Error: Combining Both

Definition and Decomposition

$$\mathrm{MSE}(\hat{\theta}) = \underbrace{\mathrm{Var}(\hat{\theta})}_{\text{Variance}} + \underbrace{[\mathrm{Bias}(\hat{\theta})]^2}_{\text{Bias}^2}$$

This is why we don't only look for unbiased estimators: a slightly biased estimator with much lower variance can have **lower MSE overall**.

Drag the slider to move along the complexity axis and read off the trade-off at that point.



Example: A Little Bias Can Pay Off

For $X_1, \ldots, X_n \sim N(\mu, 1)$, compare two estimators of μ:

- $\hat{\mu}_1 = \bar{X}$ (unbiased)
- $\hat{\mu}_2 = c\bar{X}$ for some $0 < c < 1$ (biased — "shrunk" toward 0)

Their mean squared errors are:

$$\mathrm{MSE}(\hat{\mu}_1) = \frac{1}{n} \qquad \mathrm{MSE}(\hat{\mu}_2) = \frac{c^2}{n} + (1-c)^2\mu^2$$

Takeaway

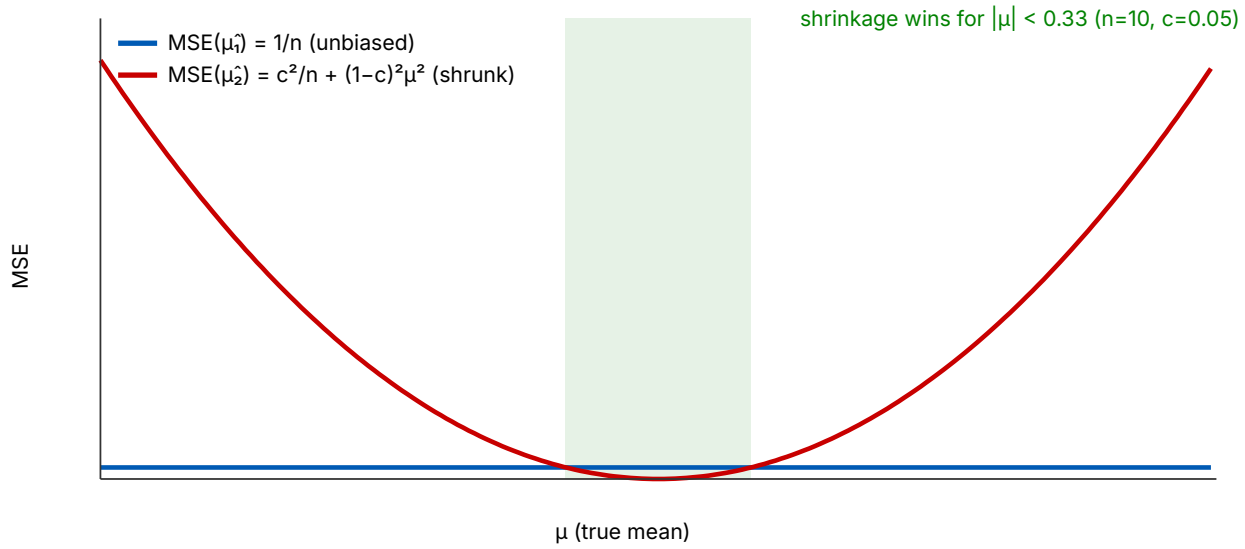
When μ is close to 0, the biased $\hat{\mu}_2$ can have **smaller** MSE than the unbiased $\bar{X}$. This is the idea behind shrinkage estimators like Ridge regression.

A Little Bias Can Pay Off, Visualized

Adjust n and c and see where the shrunk estimator beats the unbiased one (shaded region).

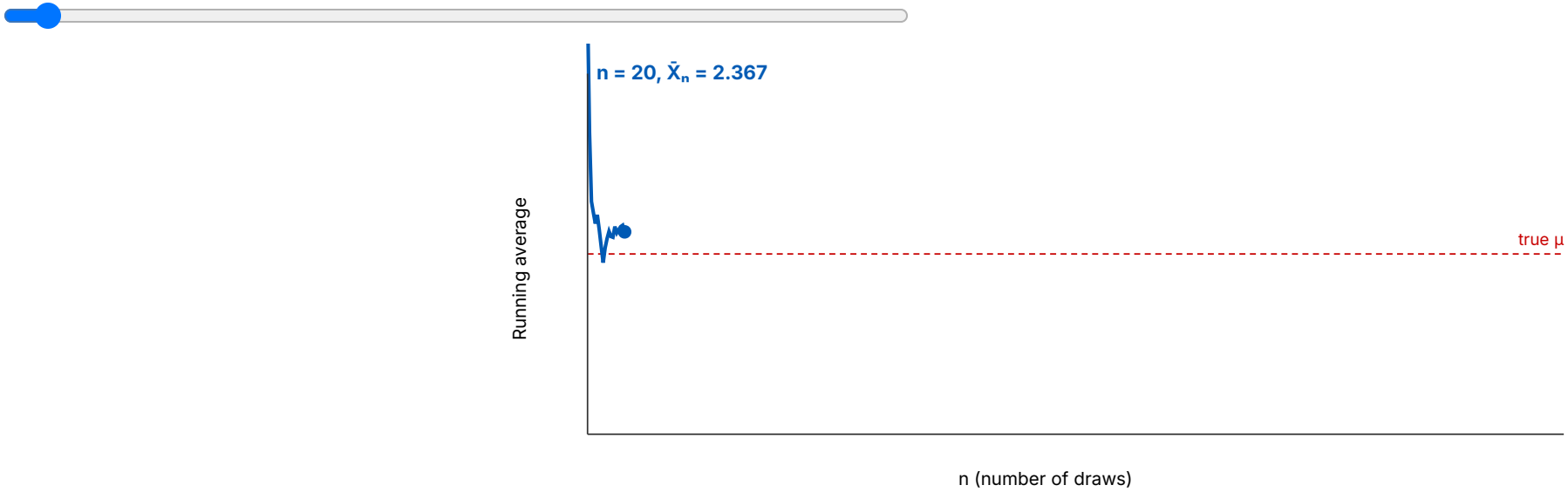
n (sample size)

c (shrinkage factor)



The Law of Large Numbers, One Path at a Time

Consistency of $\bar{X}_n$ rests on the **Law of Large Numbers**: as we add more draws, the running average settles down near the true μ. **Drag the slider** to reveal more of the path.



Consistency: Getting It Right *Eventually*

Unbiasedness is about being right *on average* for a given n. **Consistency** instead asks: does the estimator get closer and closer to the truth as we collect more data?

Definition

$\hat{\theta}_n$ is **consistent** if $\hat{\theta}_n \xrightarrow{P} \theta$ as $n \to \infty$, i.e. for every $\epsilon > 0$:

$$\lim_{n \to \infty} P\left(|\hat{\theta}_n - \theta| > \epsilon\right) = 0$$

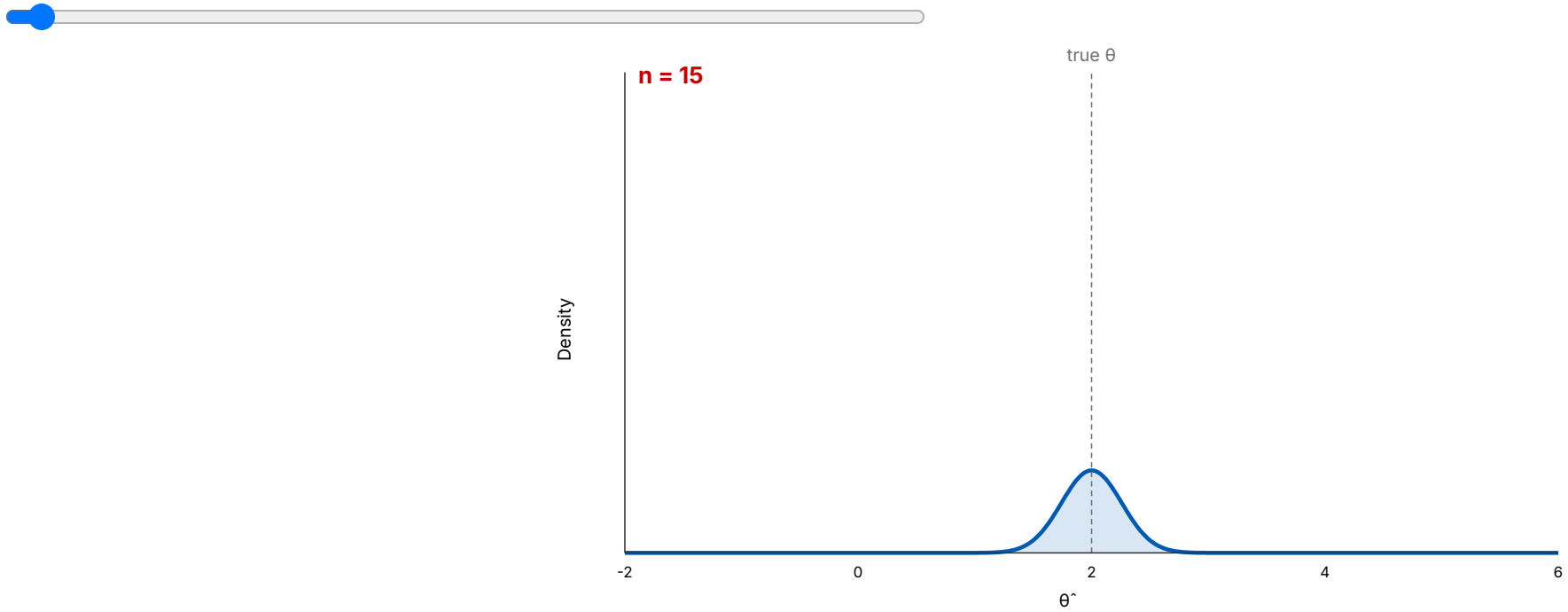
An Easy Way to Check It

If both of these hold, $\hat{\theta}_n$ is automatically consistent:

$$\lim_{n\to\infty} \text{Bias}(\hat{\theta}_n) = 0 \qquad \text{and} \qquad \lim_{n\to\infty} \text{Var}(\hat{\theta}_n) = 0$$

Consistency in Pictures

As n grows, the sampling distribution of $\hat{\theta}_n$ concentrates more and more tightly around the true value θ. **Drag the slider.**



Efficiency: How Low Can the Variance Go?

Among all unbiased estimators, some are more **precise** than others. **Efficiency** asks: is $\hat{\theta}$ the most precise unbiased estimator possible?

It turns out there is a hard floor on how small the variance of an unbiased estimator can ever be — no amount of cleverness can beat it.

That floor is the Cramér-Rao Lower Bound.

The Cramér–Rao Lower Bound

Fisher Information

$$I(\theta) = -\operatorname{E}\left[\frac{\partial^2 \ln f(X;\theta)}{\partial \theta^2}\right]$$

Roughly: how sharply peaked the likelihood is. More information means the data pins down θ more precisely.

Cramér-Rao Lower Bound (CRLB)

For *any* unbiased estimator $\hat{\theta}$ based on n observations:

$$\operatorname{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)}$$

An unbiased estimator that *achieves* this bound is called **efficient**.

Example: The Sample Mean Is Efficient

For $X_i \sim N(\mu, \sigma^2)$, estimating μ:

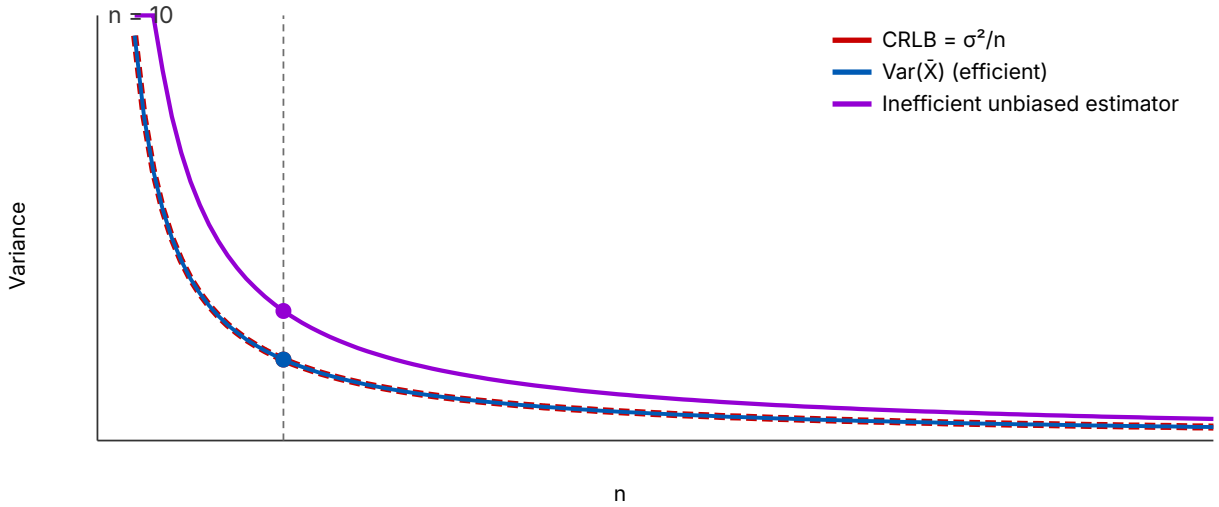
- Fisher information: $I(\mu) = 1/\sigma^2$
- Cramér-Rao bound: $\mathrm{CRLB} = \sigma^2/n$
- Actual variance: $\mathrm{Var}(\bar{X}) = \sigma^2/n$

Conclusion

$\mathrm{Var}(\bar{X})$ exactly meets the CRLB, so $\bar{X}$ is an **efficient** estimator of μ — no unbiased estimator can do better.

Efficiency, Visualized

Drag the slider to change n: notice $\mathrm{Var}(\bar{X})$ sits exactly on the CRLB floor, while a wasteful-but-unbiased estimator stays above it.



Advanced Topics

Sufficient Statistics: Throwing Away the Noise

Do we really need to keep the *entire* dataset to estimate θ, or does a short summary contain all the useful information?

Definition

A statistic $T(X)$ is **sufficient** for θ if, once you know $T(X)$, the rest of the data tells you nothing more about θ.

Example: to estimate the probability of heads p from n coin flips, the *order* of heads and tails is irrelevant — only the *total count* of heads matters. That count is a sufficient statistic.

Finding Sufficient Statistics: The Factorization Theorem

Factorization Theorem

$T(X)$ is sufficient for θ if and only if the likelihood splits as

$$f(x;\theta) = g(T(x),\theta)\cdot h(x)$$

i.e. θ only ever enters through $T(x)$.

Example: Bernoulli trials

For $X_1,\dots,X_n \sim \mathrm{Ber}(p)$, the likelihood is

$$L(p) = p^{\sum x_i}(1-p)^{n-\sum x_i}$$

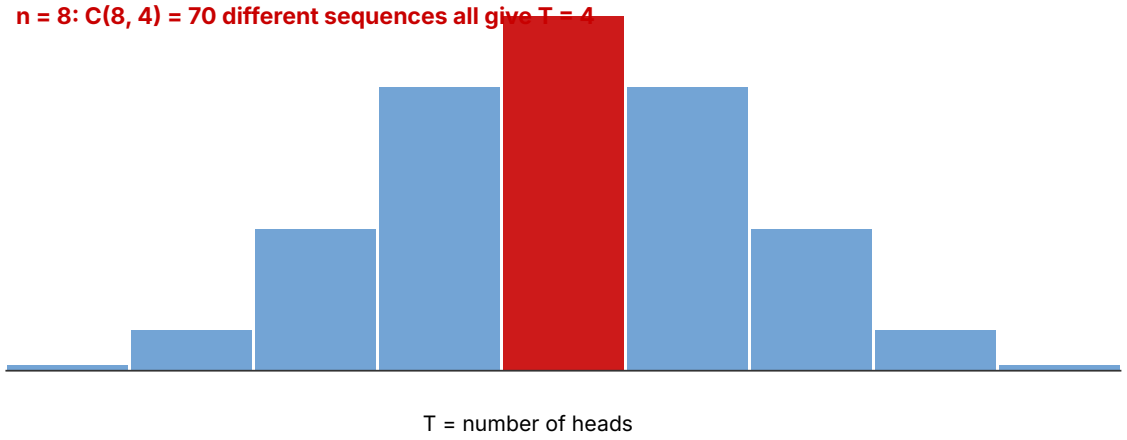
Since p only appears through $\sum x_i$, the statistic $T(X) = \sum_{i=1}^{n} X_i$ is sufficient for p.

Sufficiency, Visualized

All sequences with the same count $T = \sum x_i$ are **interchangeable** for learning p. **Adjust** n and T: the bar shows how many raw sequences collapse into that one value of T.

n (number of flips)

T (highlighted count)



The Rao–Blackwell Theorem: Never Get Worse

Idea: if you have an unbiased estimator that doesn't yet use a sufficient statistic, you can always improve it (or at worst leave it unchanged) by conditioning on that sufficient statistic.

Theorem (Rao-Blackwell)

Let $\hat{\theta}$ be an unbiased estimator of θ, and let T be sufficient for θ. Then

$$\hat{\theta}^* = \mathrm{E}[\hat{\theta} \mid T]$$

is also unbiased, and $\mathrm{Var}(\hat{\theta}^*) \leq \mathrm{Var}(\hat{\theta})$.

Rao-Blackwell in Action

For $X_1, \ldots, X_n \sim \text{Ber}(p)$:

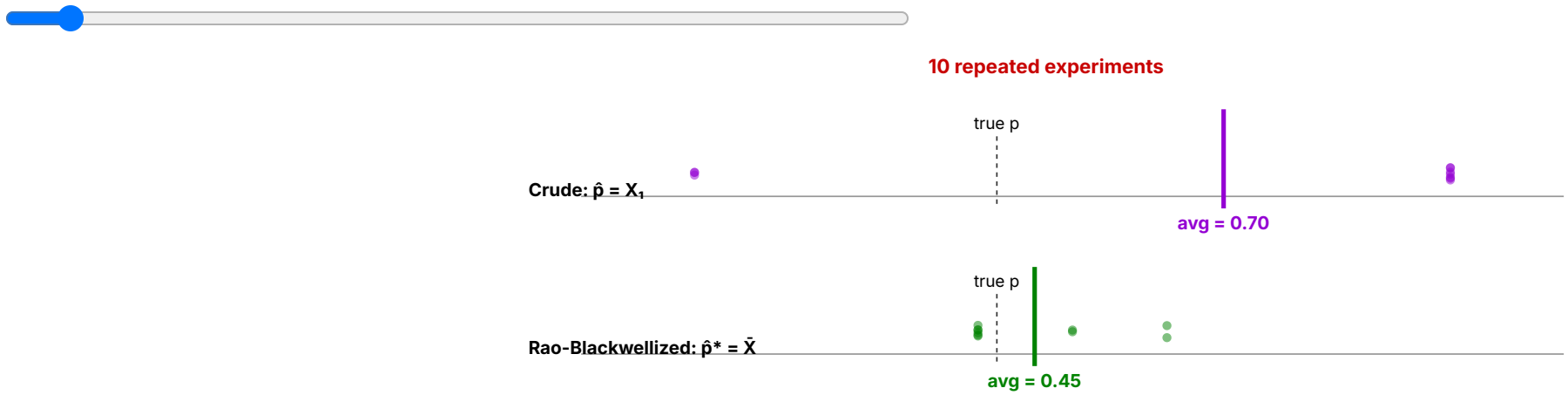
- Start with a wasteful but unbiased estimator: $\hat{p} = X_1$ (it only uses the first observation!)
- We already know $T = \sum X_i$ is sufficient for p
- Rao-Blackwellizing: $\hat{p}^* = \mathrm{E}[X_1 \mid T] = T/n$

Result

$\hat{p}^* = \bar{X}$: the sample mean, which uses *all* the data and has smaller variance than $\hat{p} = X_1$.

Rao-Blackwell, Visualized

Each dot is one repeated experiment of $n = 8$ coin flips with true $p = 0.4$. Compare the spread of the crude estimator to the Rao-Blackwellized one.



UMVUE: The Best Unbiased Estimator

Rao-Blackwell tells us that conditioning on a sufficient statistic never hurts. **UMVUE** takes this to its logical conclusion: is there an unbiased estimator that beats *every* other unbiased estimator, for *every* value of θ?

Definition

$\hat{\theta}^*$ is the **Uniformly Minimum Variance Unbiased Estimator (UMVUE)** if it is unbiased and

$$\mathrm{Var}(\hat{\theta}^*) \leq \mathrm{Var}(\hat{\theta})$$

for every other unbiased $\hat{\theta}$, and every θ.

How Do We Know We've Found It?

Theorem (Lehmann-Scheffé)

If T is a **complete** sufficient statistic and $\hat{\theta} = g(T)$ is unbiased, then $\hat{\theta}$ is automatically the UMVUE.

What does "complete" mean?

Loosely: T has no "leftover" unbiased noise in it. Formally, T is **complete** if

$$\mathrm{E}[g(T)] = 0 \ \forall \theta \quad \Longrightarrow \quad g(T) = 0 \text{ almost surely}$$

In practice: once you've found a complete sufficient statistic, *any* unbiased function of it is automatically the best possible unbiased estimator.

Finding a UMVUE: The Recipe

(1) Find a sufficient statistic T

(2) Check that T is complete

(3) Find an unbiased function of T

(4) That function *is* the UMVUE

Good to Know

- A UMVUE doesn't always exist
- MLE and UMVUE are generally *not* the same estimator
- When a UMVUE exists, it is unique

Generalized Method of Moments

From Method of Moments to GMM

Classical method of moments: write down as many equations (moment conditions) as you have parameters, then solve them exactly.

But what if you have *more* valid equations than parameters? You can't satisfy all of them exactly at once, so instead you get as close as possible to satisfying *all* of them simultaneously.

This is the idea behind GMM

- **Classical MM:** p parameters, p moment conditions (exactly identified)
- **GMM:** p parameters, $q \geq p$ moment conditions (over-identified)

The GMM Estimator

Suppose economic theory implies a set of conditions that should hold *on average* at the true parameter value:

$$\mathrm{E}[g(X_i, \theta)] = 0, \qquad g : \mathbb{R}^d \times \mathbb{R}^p \to \mathbb{R}^q$$

Their sample counterparts will rarely hit zero exactly, so we make them **as close to zero as possible**:

$$\bar{g}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(X_i, \theta)$$

GMM Estimator

$$\hat{\theta}_{GMM} = \arg\min_{\theta} \; \bar{g}_n(\theta)' W_n \bar{g}_n(\theta)$$

where W_n is a $q \times q$ weight matrix that decides how much each condition counts.

Choosing the Weights

Not all moment conditions are equally reliable — noisier ones should count for less.

Two-Step Procedure

(1) Start with equal weights ($W = I$) to get a first-pass estimate $\tilde{\theta}$

(2) Use $\tilde{\theta}$ to see how noisy and correlated the moment conditions are

(3) Re-estimate using weights that downweight the noisier conditions

With the optimal weight matrix, GMM achieves the **smallest possible asymptotic variance** among all choices of W — the same spirit as the Cramér-Rao bound, for a much broader class of estimators.

For the curious: that smallest variance takes the compact form $V = (G'\Sigma^{-1}G)^{-1}$ — you'll see it derived in more advanced courses.

Example: Instrumental Variables Regression

Model: $y = X\beta + \epsilon$, but $\mathrm{E}[X'\epsilon] \neq 0$ — X is **endogenous**, so OLS is biased.

Fix: find instruments Z that are correlated with X but uncorrelated with ϵ:

$$\mathrm{E}[Z'\epsilon] = 0 \quad \Longrightarrow \quad g(y, X, Z; \beta) = Z'(y - X\beta)$$

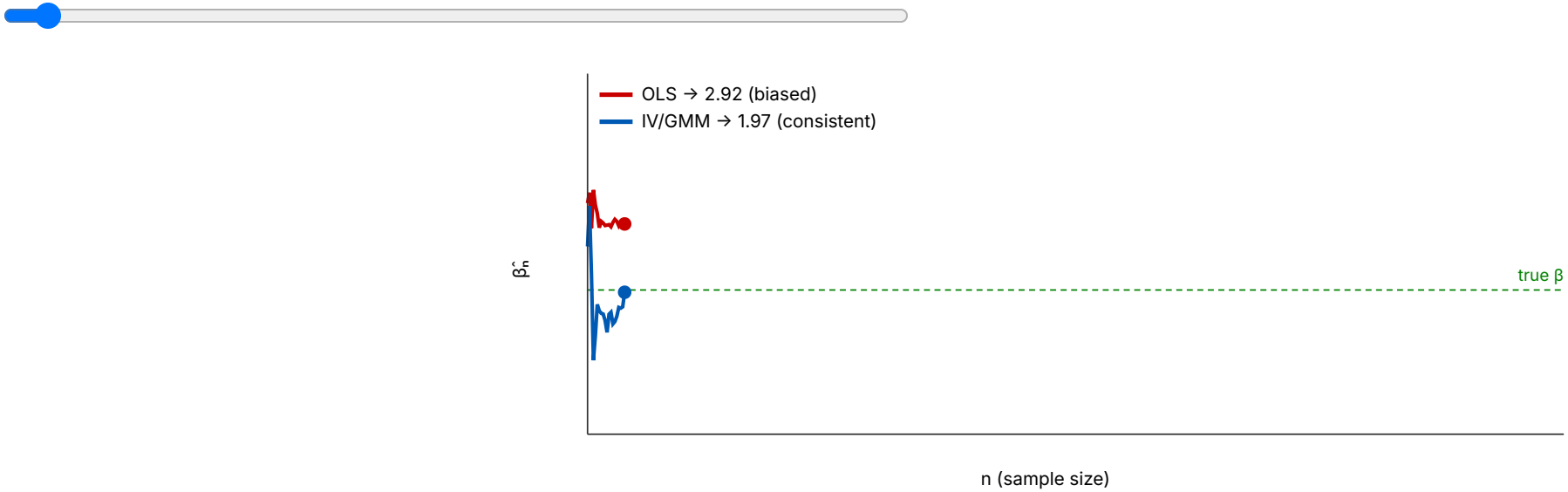
GMM / IV Estimator

$$\hat{\beta}_{GMM} = (X'ZW_nZ'X)^{-1}X'ZW_nZ'y$$

As many instruments as regressors: this is the classic IV estimator. More instruments than regressors: it's 2SLS with optimal weights.

OLS Bias vs. IV Consistency, Visualized

Each line is the running estimate as the sample grows. **Drag the slider** to add more observations and watch which estimator finds the true β.



GMM Shows Up All Over Economics

- **Instrumental variables & 2SLS:** fixing endogeneity in regression (previous slide)
- **Rational expectations models:** Euler equations linking today's consumption choices to future returns
- **Asset pricing (CAPM):** relating an asset's expected return to its exposure to market risk
- In each case, economic theory supplies the moment conditions — GMM turns them into an estimator, without requiring a full distributional assumption

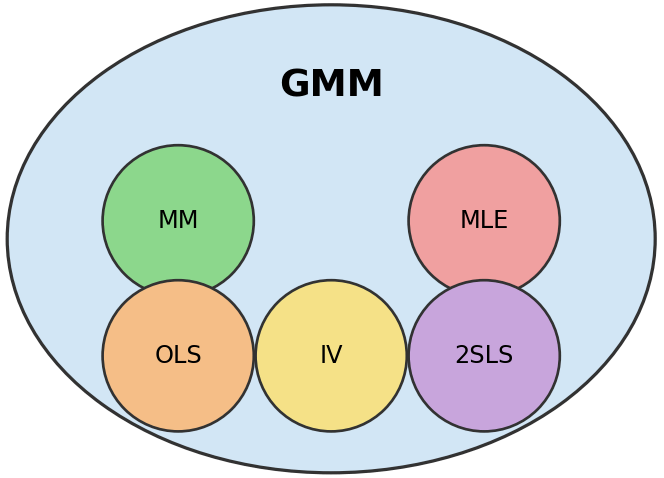
Model Testing and GMM's Family Tree

J-Test for Over-Identification

When $q > p$, the extra moment conditions can be used to test the model itself:

$$J = n \cdot \bar{g}_n(\hat{\theta})' W \bar{g}_n(\hat{\theta}) \xrightarrow{d} \chi^2_{q-p}$$

A large J suggests some moment condition — and so the model — may be wrong.



Key Insight

Many familiar estimators are just GMM with a particular choice of moment conditions and weights.

When (Not) to Reach for GMM

Good fit when…

- You have more moment conditions than parameters
- There's an endogeneity problem
- The model is nonlinear
- You don't want to assume a full distribution

Watch out for…

- GMM needs fairly large samples to behave well
- Results can be sensitive to *which* moments you pick

Comparison of Methods

Comparing the Estimation Methods

Property	MM	GMM	LS	MLE	UMVUE
Always exists	✅	✅	✅	✅	❌
Unbiased	❌	❌	✅[1]	❌	✅
Consistent	✅ (usually)	✅	✅	✅	✅
Asymptotically efficient	❌	❌	✅[2]	✅	✅
Needs a distribution	❌	❌	❌	✅	✅

[1] Linear models [2] Normal errors

Rules of Thumb

- **Quick and simple:** Method of Moments
- **Know the distribution:** MLE
- **Linear, no distribution needed:** Least Squares
- **Endogeneity or extra moment conditions:** GMM
- **Need unbiasedness in a small sample:** UMVUE

Key Takeaways

(1) **Good estimators balance bias and variance** — that's what MSE measures

(2) **Consistency** means more data brings $\hat{\theta}_n$ closer to the truth

(3) **Efficiency** and the Cramér-Rao bound tell us how precise an unbiased estimator can possibly be

(4) **Sufficient statistics** let us summarize the data without losing information about θ

(5) **Rao-Blackwell** and **completeness** lead us to the **UMVUE**, the best possible unbiased estimator

(6) **GMM** extends the method of moments to handle extra information, endogeneity, and weaker distributional assumptions

(7) **No single method wins every time** — the right choice depends on your assumptions, sample size, and goals

Beyond Point Estimation: What's Next

So far we've focused on finding a single "best guess" $\hat{\theta}$ — but that's only part of the story.

Interval Estimation

- Confidence intervals
- Prediction intervals
- Bootstrap intervals

Bayesian Methods

- Priors and posteriors
- Credible intervals
- Hierarchical models

Robust Estimation

- M-, L-, and R-estimators
- Influence functions

Modern Extensions

- High-dimensional estimation
- Shrinkage (LASSO, Ridge)
- Machine learning & causal inference