Estimation

From a Point Estimate to an Interval

You Only Ever See One Sample

You compute $\hat{\theta}$ from the one dataset you have. A colleague with a different sample from the same population gets a different number. Neither of you is wrong.

The object this whole deck is about

The **sampling distribution** of $\hat{\theta}$: the distribution of the numbers you would get by repeating the study. Every idea that follows is a statement about one of its features.

Where it sits

Bias, and what it costs — the first half of this deck

How wide it is

Variance, its floor, and the **interval** that reports it — the second half

A point estimate is one draw from that distribution. An interval reports its width.

Two Roads to the Same Estimator

Big picture: least squares and maximum likelihood both provide answers to the same question — *which parameter values fit 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

Key Insight

Under the right assumptions, these two different ideas lead to the **exact same estimator**. The next slide shows 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 equivalent to* maximum likelihood.

Where the Distribution Sits

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

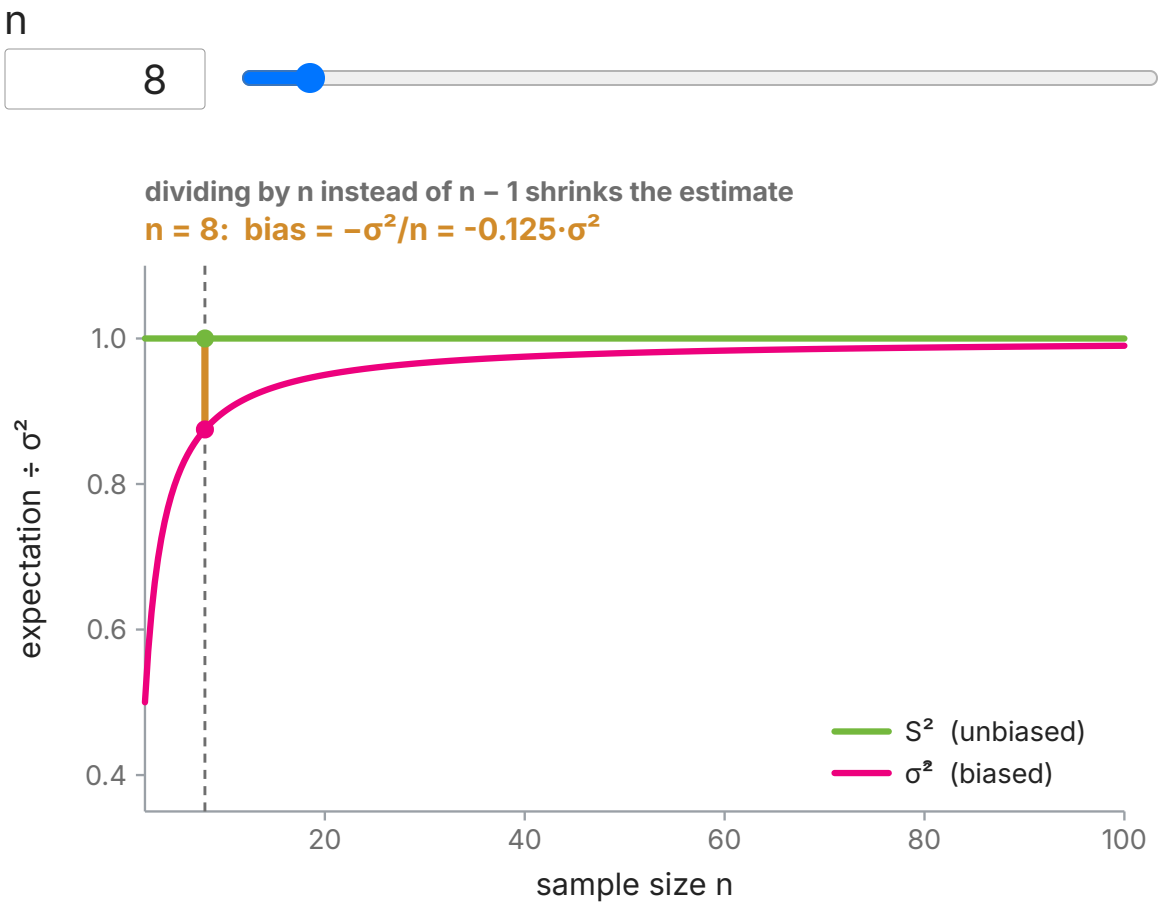
Unbiased

- $\bar{X}$ 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

$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2$ — the bias vanishes as n grows, which is the first hint that *unbiased* and *consistent* are different promises.



Unbiasedness, Visualized

repeated samples

150

sample size n

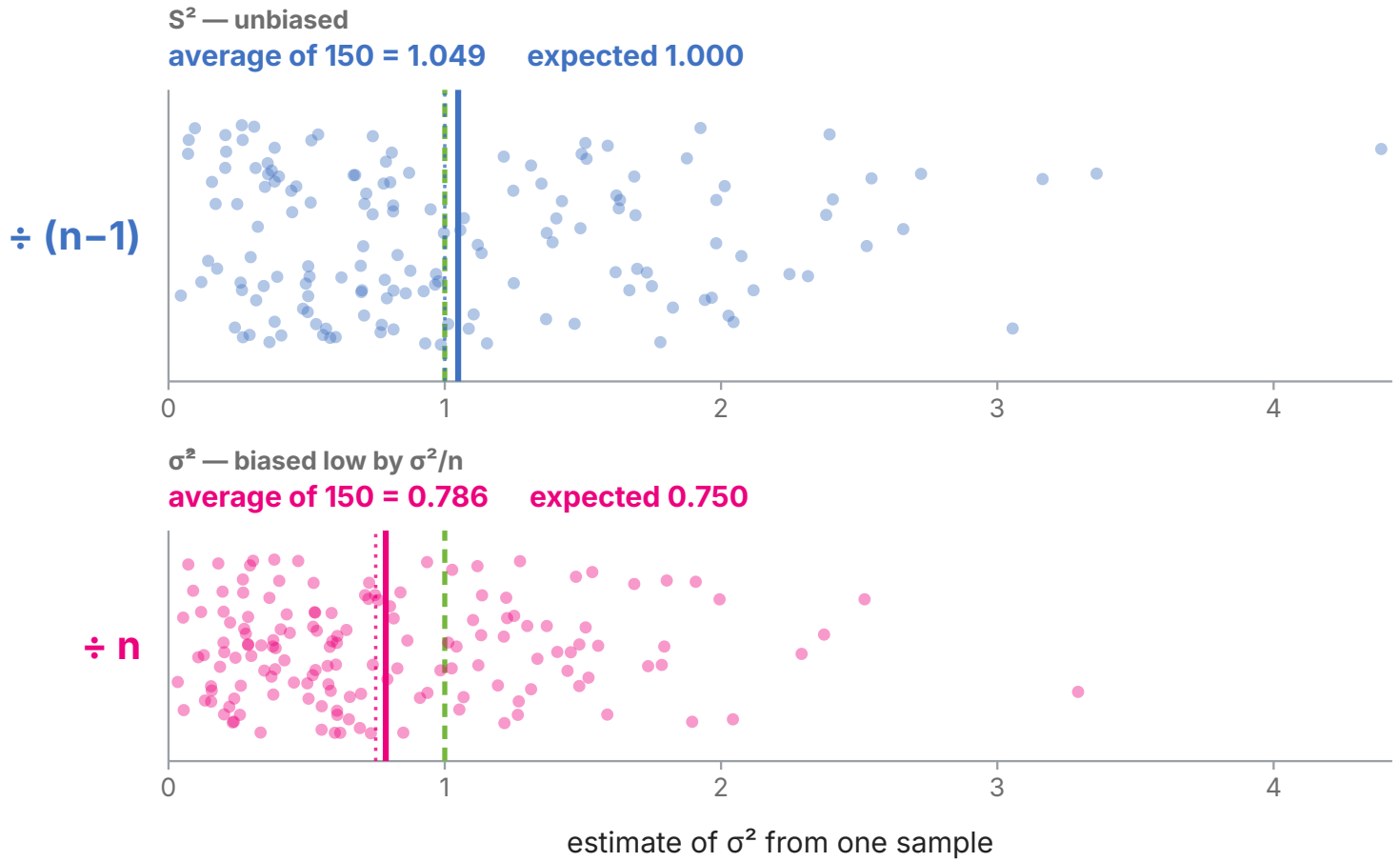
4

New samples

Every sample is drawn honestly from the true distribution — nothing is shifted. Both rows are handed the **same** data and differ only in the divisor.

The dotted rule is where each estimator is *expected* to land. Add repetitions and each solid average walks onto its own dotted rule — one of which is not σ^2.

Pull the repetitions down and the averages become unreliable: unbiasedness is a statement about the long run, not about one study.



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

shots per target

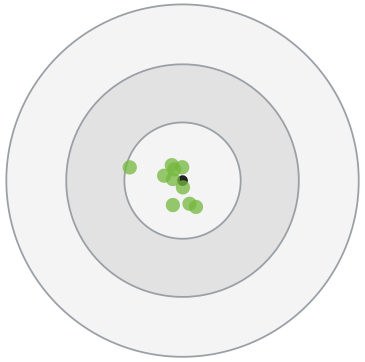
10

Rows differ in **bias** — whether the cloud is centred on the bullseye. Columns differ in **variance** — how tightly it is packed.

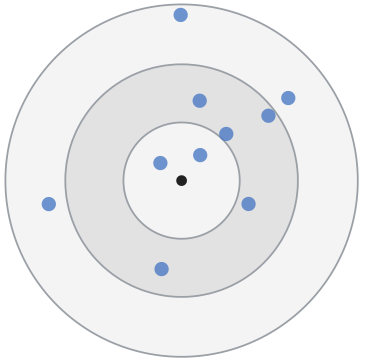
At one or two shots the four targets are indistinguishable. The pattern is a property of the *distribution*, not of any single estimate you will ever compute.

Keep in mind: each dot is an estimate, not a data point!

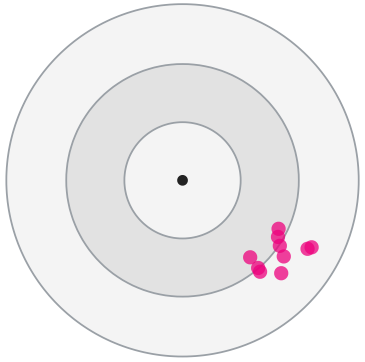
low bias, low variance



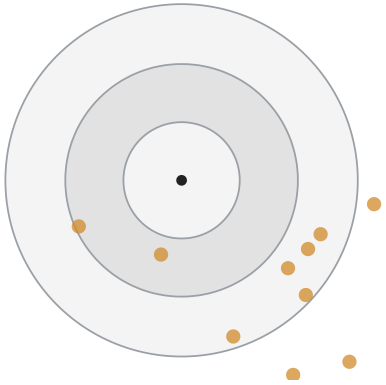
low bias, high variance



high bias, low variance



high bias, high variance



Mean Squared Error: Combining Both

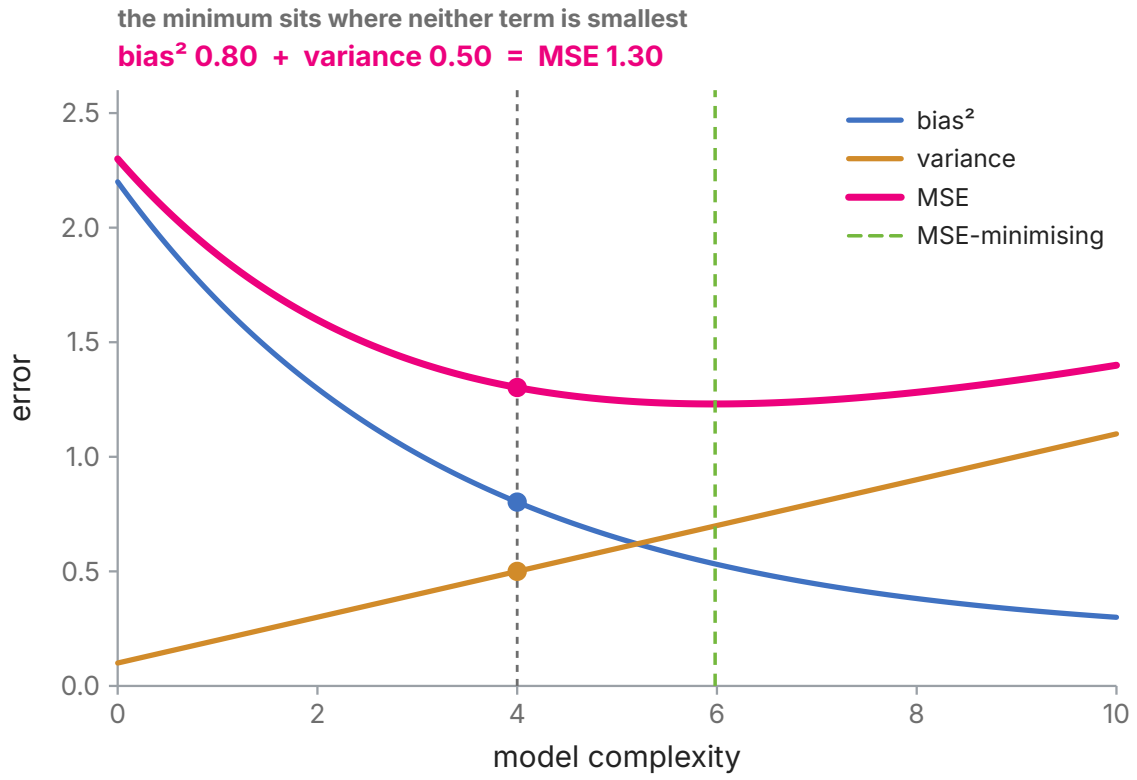
Definition and Decomposition

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

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 read the trade-off at each point.

model complexity

4



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

n

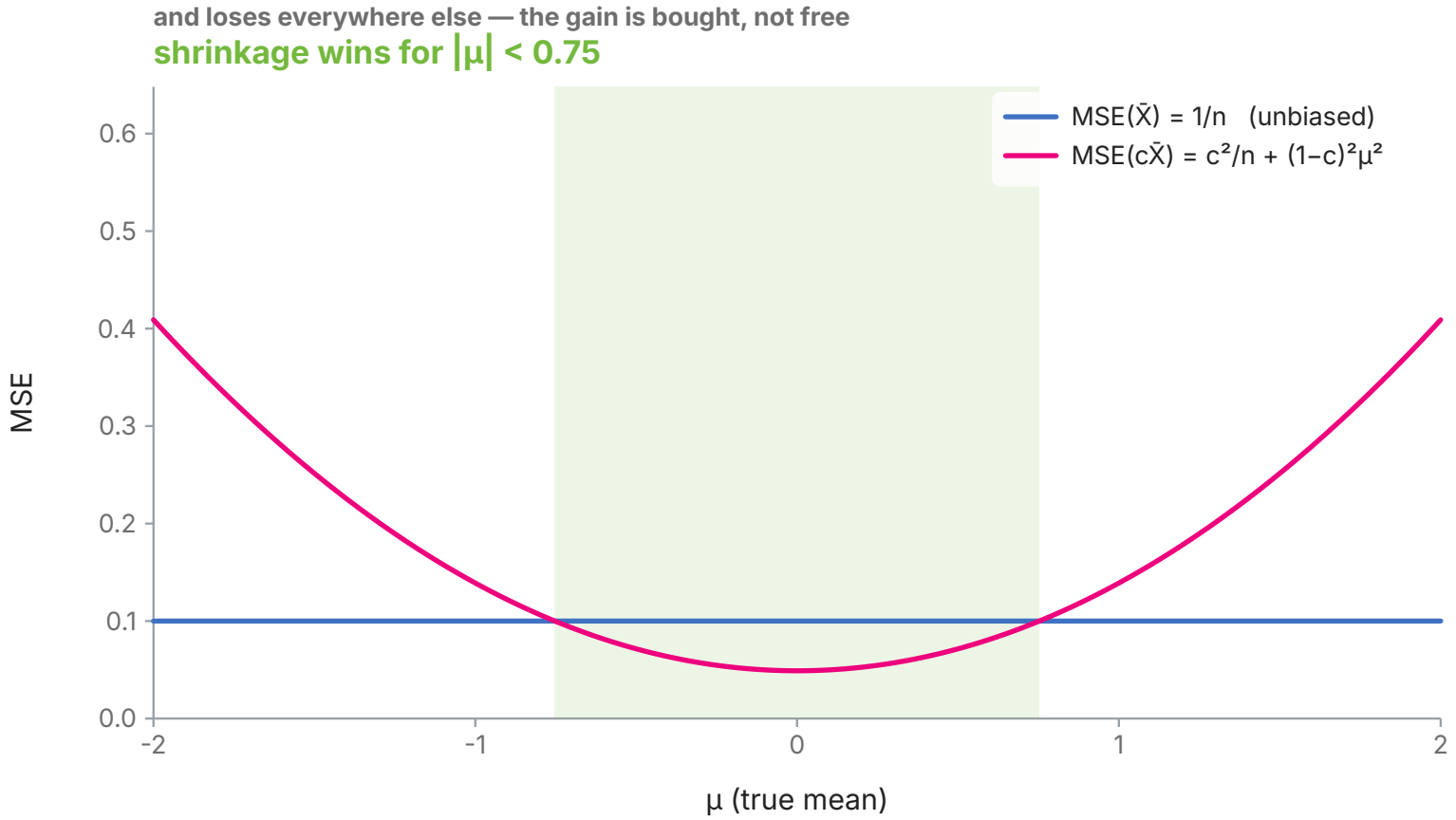
10

c

0.7

Shaded: where the shrunk estimator has the **lower** MSE.

Shrinking hard (c small) wins big near $\mu = 0$ and loses badly far from it. Raising n narrows the shaded band — with enough data there is little left to gain by biasing the estimate.



What More Data Buys

The Law of Large Numbers, One Path at a Time

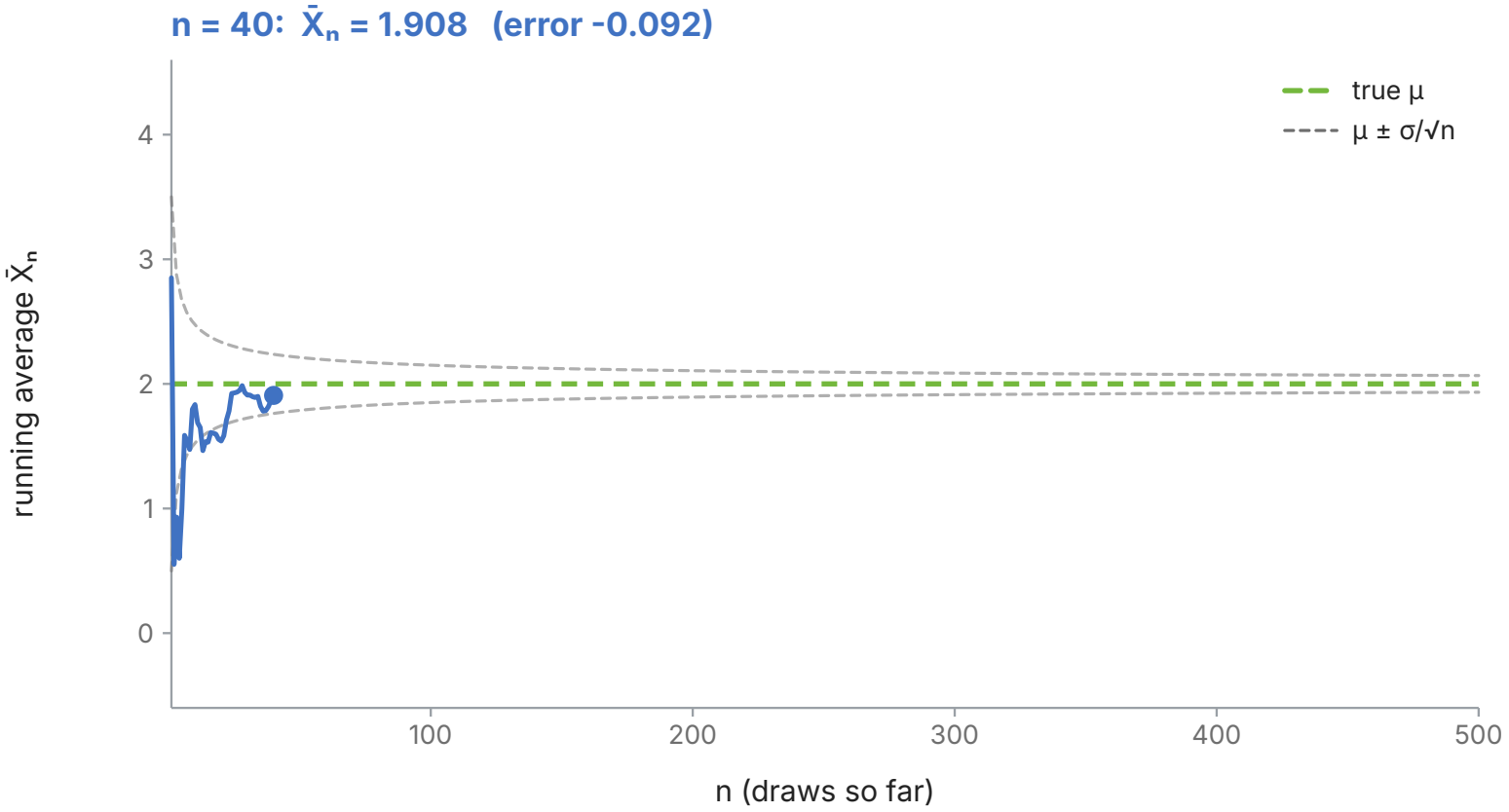
draws n

40

New path

Every property so far was stated for a *fixed* n. Now let n grow: the running average settles down near the true μ.

The dashed funnel is $\mu \pm \sigma/\sqrt{n}$ — the path is squeezed by arithmetic, not by luck. Hold on to that shape: by the end of this deck it will be an interval, drawn the other way round.



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

n

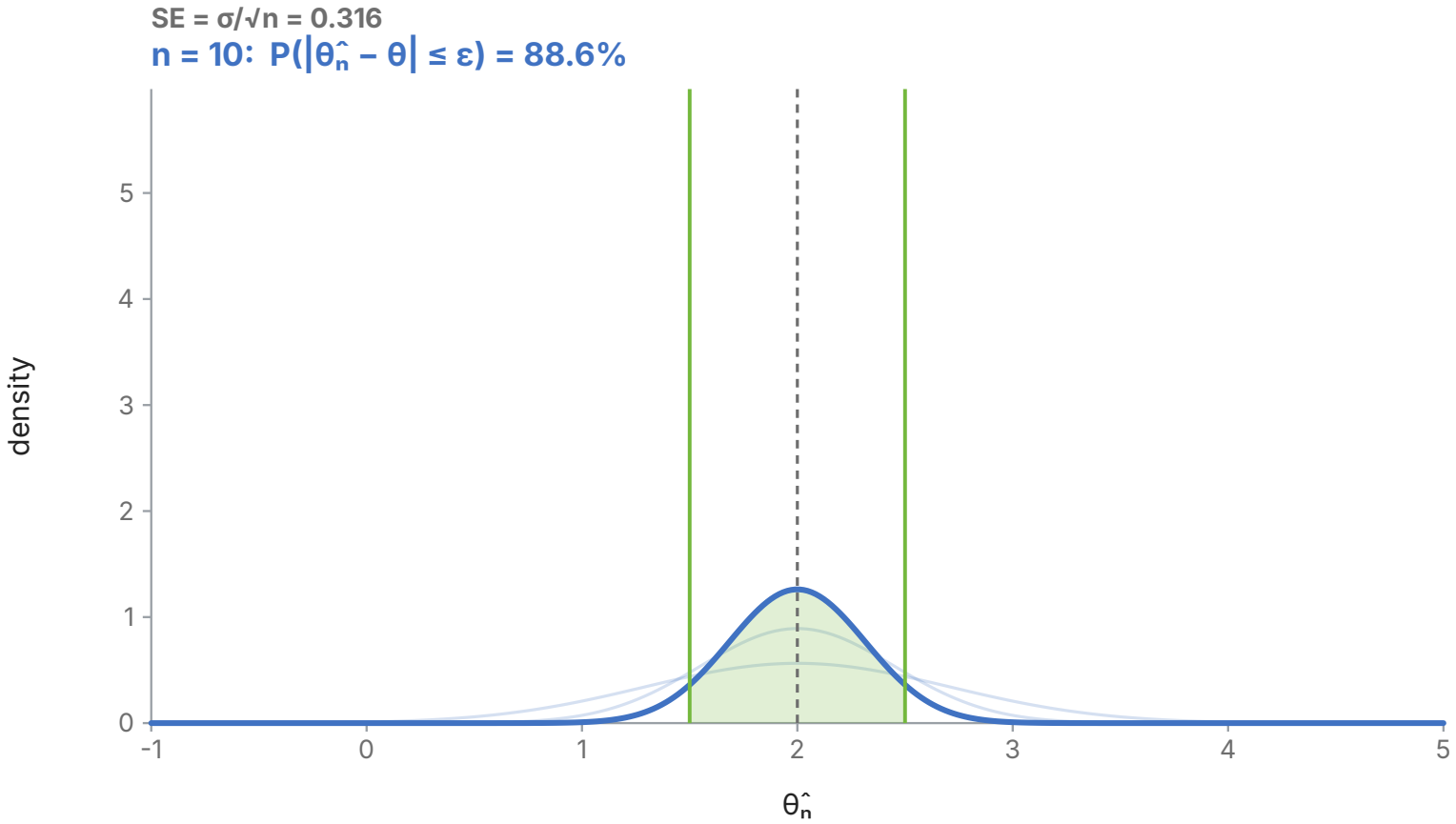
10

ε

0.5

The whole sampling distribution collapses onto θ; the faint curves are the smaller sample sizes this one grew out of.

Consistency is the shaded percentage going to 100 for **every** ϵ, however small. It says nothing about how fast — that is the next section.



How Precise Can We Get?

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 given some sample size?

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

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

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:

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

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: $\mathbf{Var}(\bar{X}) = \sigma^2/n$

Conclusion

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

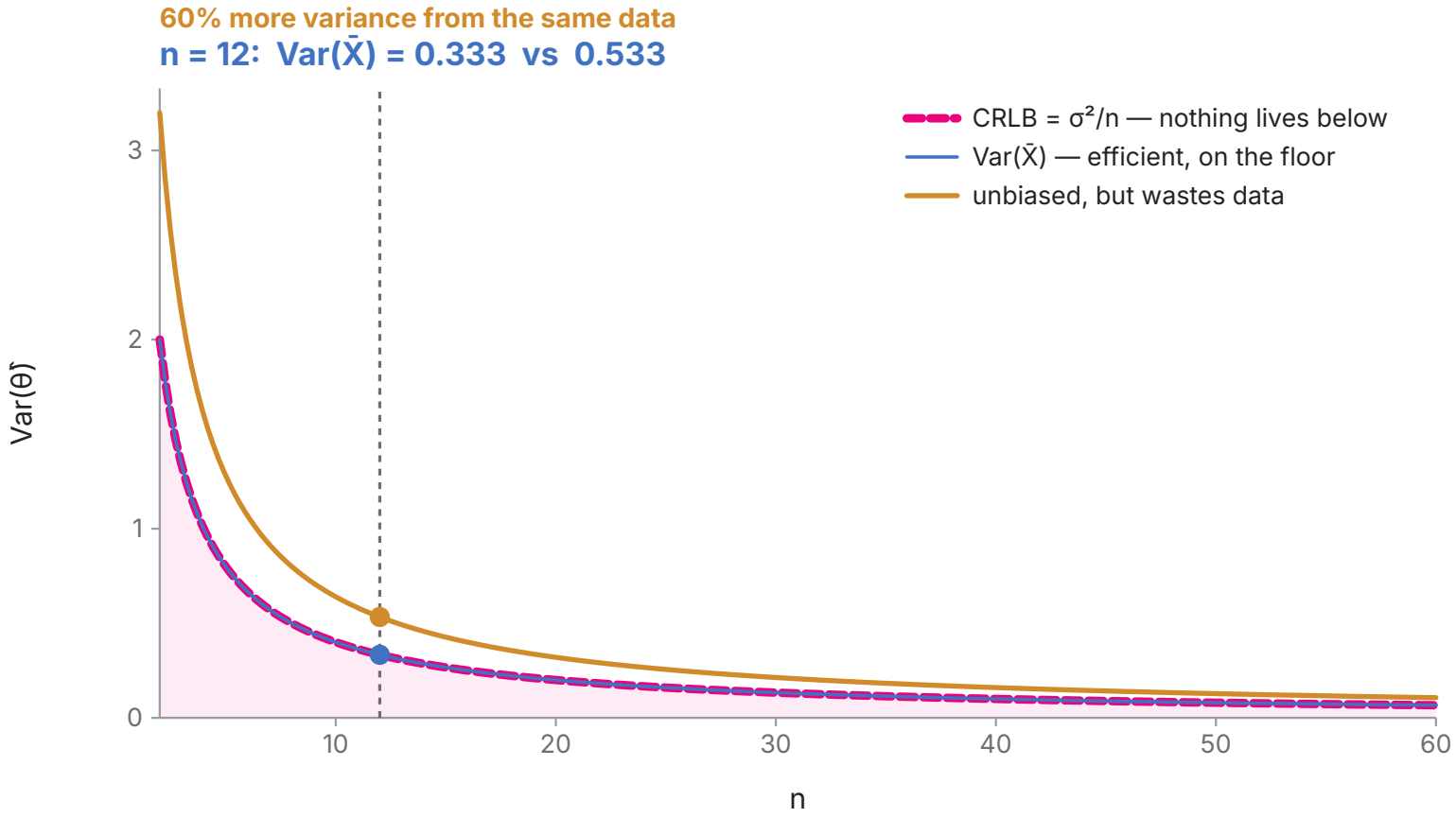
Efficiency, Visualized

n

12

$\mathrm{Var}(\bar{X})$ sits exactly on the CRLB floor; the shaded region below it is empty for **every** unbiased estimator, at every n.

The wasteful estimator is unbiased too — it simply throws away information. It never crosses the floor, it just needs more data to reach the same precision.



The Bound Has Two Consequences

The CRLB is the last statement this deck makes about $\mathrm{Var}(\hat{\theta})$ in the abstract. It splits the rest of the course in two.

Can the floor be reached?

If a floor exists, is there a recipe that lands on it? That is **sufficiency**, **Rao-Blackwell** and the **UMVUE** — picked up again near the end of this deck.

What is the floor *for*?

A variance nobody reports is useless. Turning $\mathrm{Var}(\hat{\theta})$ into a number on a page is **interval estimation** — everything from here to the bootstrap.

We take the second road first, because it is the one that reaches a published table.

From Spread to Interval

That Variance Has a Name

Every slide of the last section computed $\mathbf{Var}(\bar{X}) = \sigma^2/n$. Its square root is the number you report.

Standard deviation

$$\mathrm{SD}(X_i) = \sqrt{\mathbf{Var}(X_i)}$$

A property of the **population**. It does not shrink with n; a larger sample only estimates it better.

Standard error

$$\mathrm{SE}(T) = \sqrt{\widehat{\mathbf{Var}}(T)}$$

A property of the **estimator** $T = T(X_1, \ldots, X_n)$ — the width of the sampling distribution we have been drawing all along.

Every interval from here is built from the second one

$\bar{x} \pm 1.96\, s$ says where the *data* lie; $\bar{x} \pm 1.96\, s/\sqrt{n}$ says where the *mean* lies. The Cramér-Rao bound is a floor on the second, and so on every interval below.

From a Point to a Range

A point estimate answers *what is our best guess*. It never answers *how good is the guess* — and the second question is the one a policy note has to survive.

The object of the next four sections

An **interval estimator** is a pair of statistics $\hat{\theta}_L(X) \leq \hat{\theta}_U(X)$, and its **coverage probability** is

$$P_\theta\big(\hat{\theta}_L \leq \theta \leq \hat{\theta}_U\big).$$

If this equals $1 - \alpha$ for every θ, the interval has **confidence level** $1 - \alpha$.

Read the probability carefully

$\hat{\theta}_L$ and $\hat{\theta}_U$ are the random variables. θ is a fixed unknown constant. The probability is over **intervals**, not over θ.

Coverage, Live

Each horizontal line is one 95% interval from one sample — the same repeated sampling as the dot strips earlier, with a bar instead of a dot. The vertical rule is the true μ, which in real life you never see.

Sample size n

20

Level

95%

Population

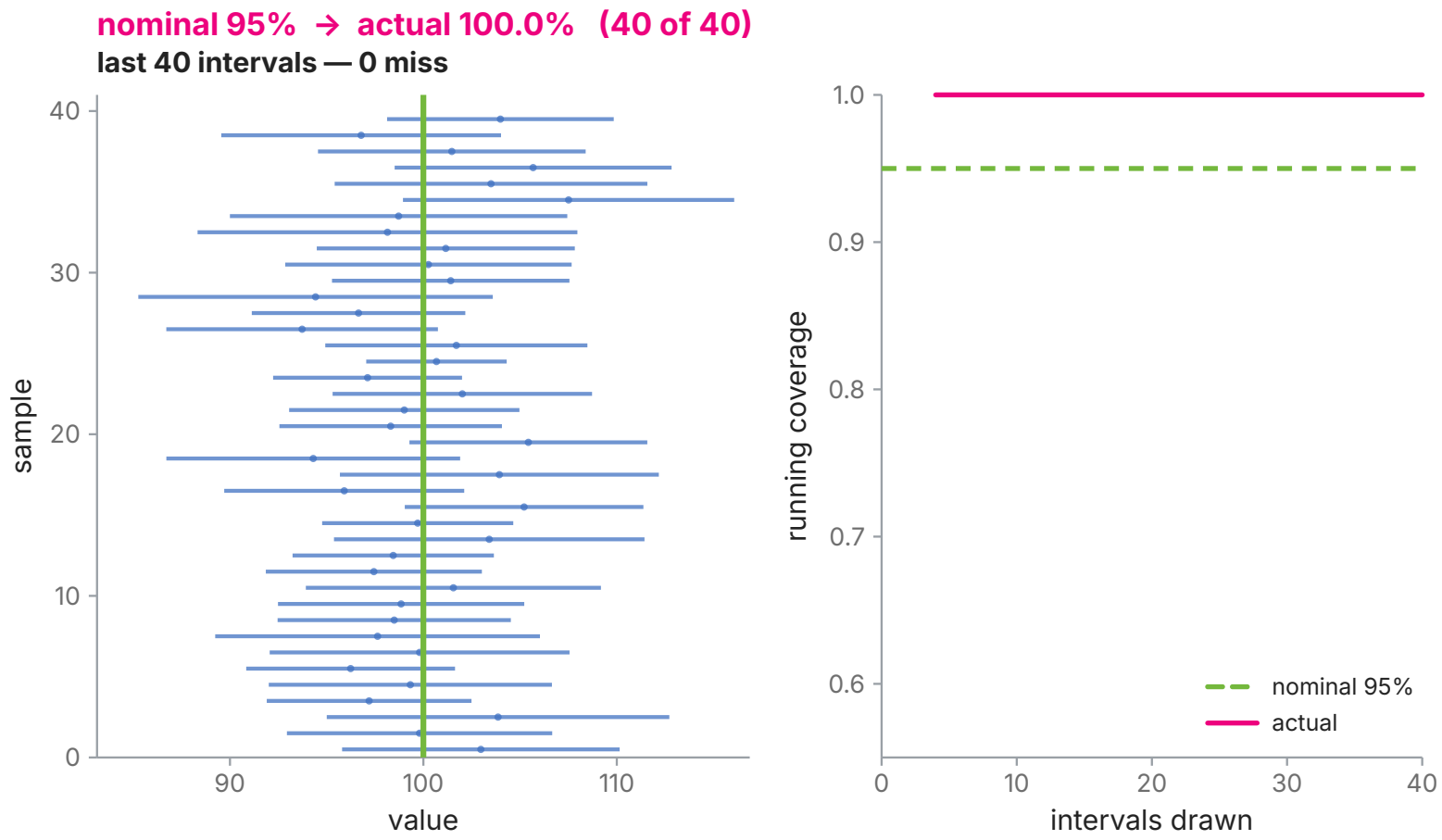
Normal

Method

t interval (correct)

Draw 40 more

95% is a promise about the **procedure**, not about the interval on your screen. Break an assumption — a skewed population at small n, a rare binary event, or z where t belongs — and the promise quietly stops being kept.



What the Interval Plot Shows

- With the t interval on a normal population, the running coverage settles on the nominal level — the procedure keeps its promise
- A **skewed** population at $n = 10$ under-covers: the interval is symmetric, the sampling distribution is not
- **Bernoulli with $p = 0.05$** is worse still, and the intervals visibly extend below zero
- Substituting z for t at small n under-covers — which is precisely why t exists

The point

Nominal coverage is a property of the **procedure under its assumptions**. The label keeps saying 95% also if the assumptions are violated.

The scatter of the dots in *Unbiasedness, Visualized* and the length of these bars are the same quantity: one standard error, times a critical value.

What a Confidence Interval Is *Not*

Correct interpretation

- 95% of intervals **built this way** contain θ
- The procedure fails 5% of the time, in the long run
- Wider interval $\Rightarrow$ less precise estimate

A confidence interval is NOT

- $P(\theta \in [\hat{\theta}_L, \hat{\theta}_U] \mid \text{data}) = 0.95$
- A range containing 95% of the **data**
- A range containing 95% of **future estimates**
- "They overlap, so the difference is not significant"

Once you have seen the data, the interval either contains θ or it does not. Nothing random is left to attach a probability to.

The third item is a genuinely different number: about 83% of future estimates fall inside a given 95% interval.

Building Intervals

Pivotal Quantities

Definition — Pivotal Quantity

A **pivot** is a function $Q(X, \theta)$ of the data and the parameter whose distribution does **not** depend on θ.

Find a pivot, look up its quantiles, and invert:

$$P\left(q_{\alpha/2} \le Q(X,\theta) \le q_{1-\alpha/2}\right) = 1-\alpha \quad \Longrightarrow \quad \text{solve for } \theta$$

Known variance

$$Q = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

$$\bar{x} \pm z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}}$$

Unknown variance

$$Q = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}$$

$$\bar{x} \pm t_{n-1,\,1-\alpha/2}\,\frac{s}{\sqrt{n}}$$

An Interval That Is Not Symmetric

The $\pm$ form is not the definition of an interval — it is what a **symmetric** pivot gives you.

Pivot for a normal variance

$$Q = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

Inverting $P\left(\chi^2_{n-1,\alpha/2} \le Q \le \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha$ gives

$$\left[\ \frac{(n-1)s^2}{\chi^2_{n-1,\,1-\alpha/2}}\ ,\ \frac{(n-1)s^2}{\chi^2_{n-1,\,\alpha/2}}\ \right]$$

Note what changed

The χ^2 distribution is skewed, so the two critical values are not symmetric about $n-1$ — and s^2 is **not** at the centre of its own interval.

The Wald Interval for a Proportion

The interval in every textbook, $\hat{p} \pm z\sqrt{\hat{p}(1-\hat{p})/n}$, applied to unemployment rates, vote shares and take-up rates.

n

40

Interval

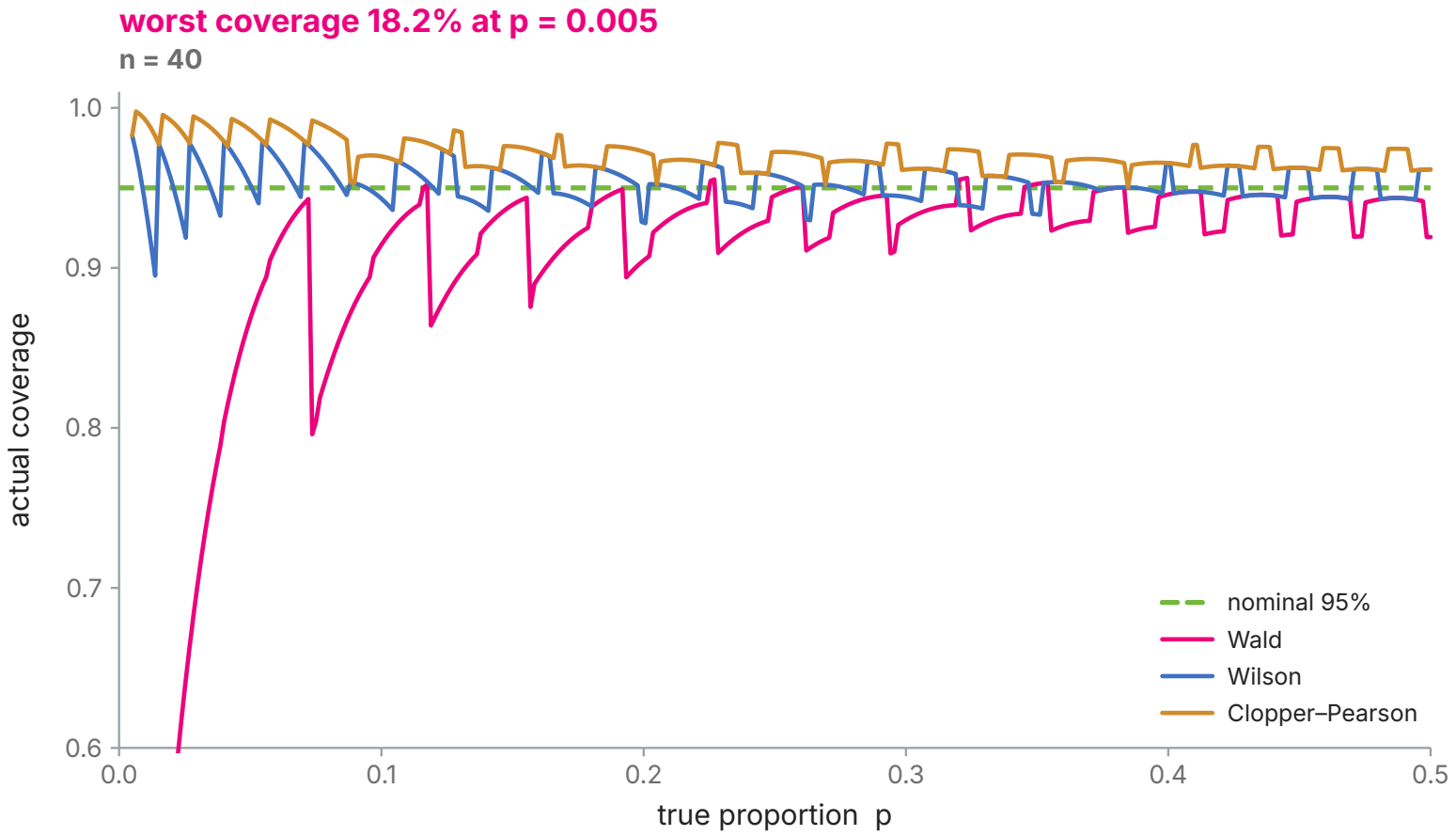
All three

Level

95%

Coverage here is **exact**, not simulated: for each p we sum $\binom{n}{k}p^k(1-p)^{n-k}$ over every k whose interval covers p.

Raising n does not smooth the sawtooth away. It moves it.



What the Sawtooth Means

- Coverage is a **step function** of p: the data are discrete, so the interval jumps as k does
- The Wald interval dips far below its nominal level at perfectly ordinary values of p — and raising n relocates the dips rather than removing them
- **Wilson** stays close to nominal almost everywhere, for the same computational cost
- **Clopper–Pearson** never falls below nominal — and pays for it by sitting *above*, which means intervals wider than they need to be

Conservative is not free

"Guaranteed at least 95%" and "exactly 95%" are different promises. The first costs precision on every single application.

Precision Has a Price

Half-Width and Sample Size

For the mean, the half-width of a $1-\alpha$ interval is

$$w = z_{1-\alpha/2}\,\frac{\sigma}{\sqrt{n}} \qquad \Longleftrightarrow \qquad n = \left(\frac{z_{1-\alpha/2}\,\sigma}{w}\right)^2$$

Precision is bought at a quadratic price

Halving the width costs **four times** the sample. Quartering it costs **sixteen times**.

Read forwards, this is a diagnosis: here is how precise we were.

Read backwards, it is a **design tool**: here is the sample we must buy.

This is the Cramér-Rao bound, in the currency of a budget

The floor on $\mathrm{SE}(\hat{\theta})$ is a floor on w, so the CRLB sets the **cheapest possible study** that can reach a stated precision. An efficient estimator is one that does not make you buy observations twice.

The Precision Planner, Live

$\log_{10}$ n

3.08

σ

0.5

Level

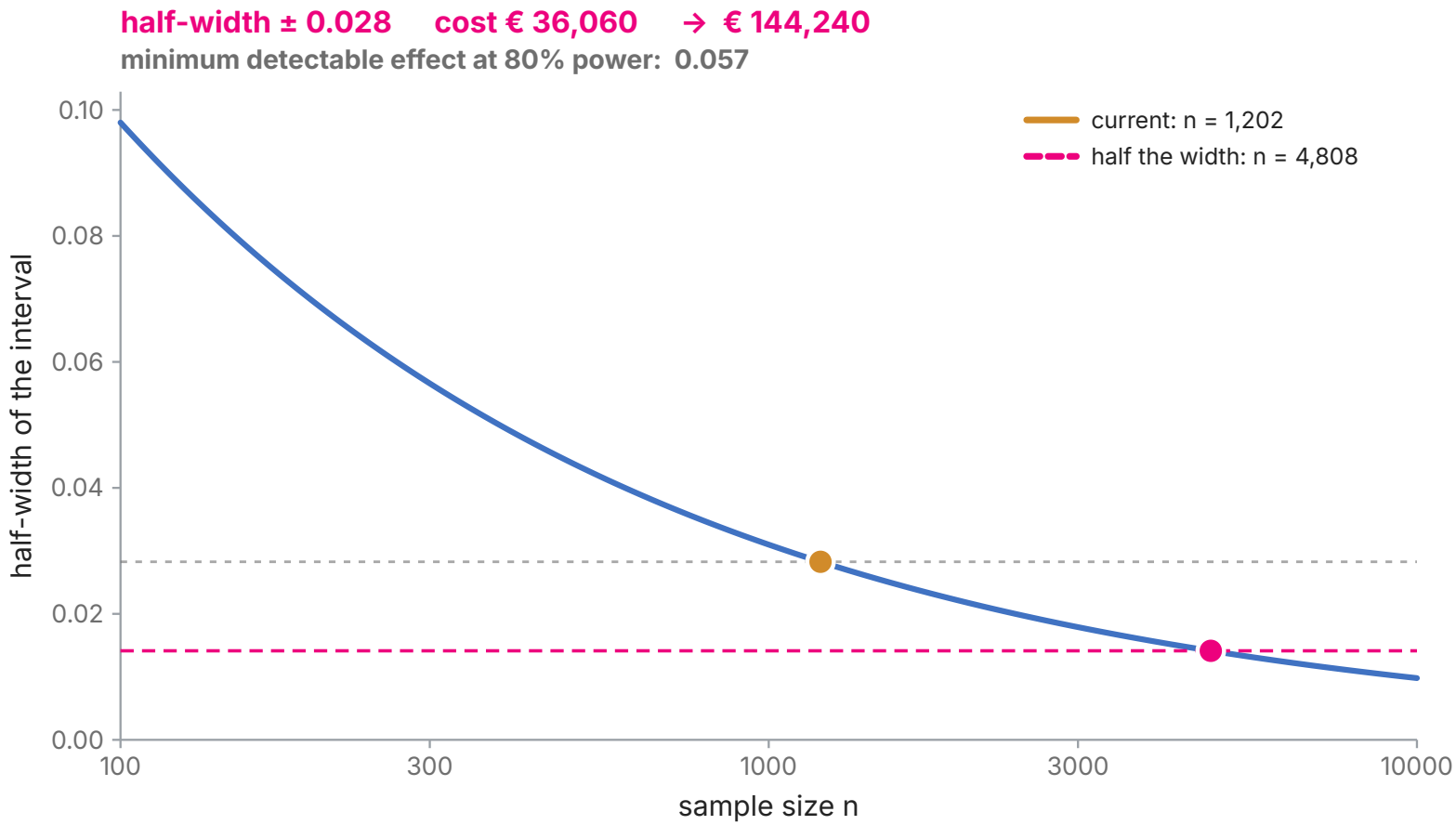
95%

€ per observation

30

The curve is $w = z\sigma/\sqrt{n}$. The marked point is your current design; the dashed line is what it would take to halve the width.

The same algebra answers the power question of the next lecture, which is why design and precision are one calculation.



Confidence Interval $\neq$ Prediction Interval

Two intervals that look identical on a chart and answer different questions.

Confidence interval — for the mean

$$\bar{x} \pm z\,\frac{\sigma}{\sqrt{n}} \xrightarrow[n\to\infty]{} \{\mu\}$$

Width $\to 0$. With enough data you know μ exactly.

Prediction interval — for the next draw

$$\bar{x} \pm z\,\sigma\sqrt{1+\tfrac{1}{n}} \xrightarrow[n\to\infty]{} \mu \pm z\sigma$$

Width $\to 2z\sigma$. It **stops**.

The distinction economists get wrong

Forecast fan charts are prediction intervals. Reading one as a confidence interval understates forecast uncertainty by a factor that grows with $\sqrt{n}$. The left-hand limit *is* the **consistency** picture from earlier; the right-hand one cannot collapse, because the next observation is not an estimator of anything.

Duality with Testing

A Confidence Interval Is an Inverted Test

Hypothesis testing, the next lecture, builds rejection regions; this deck builds intervals. They are the same object, seen from two sides.

Duality

$$\mathrm{CI}_{1-\alpha}(x) \;=\; \{\, \theta_0 \;:\; \text{the level-}\alpha \text{ test of } H_0\colon \theta = \theta_0 \text{ does not reject} \,\}$$

Equivalently, writing $p(\theta_0)$ for the p-value of the test of $H_0 : \theta = \theta_0$,

$$\mathrm{CI}_{1-\alpha}(x) \;=\; \{\, \theta_0 : p(\theta_0) > \alpha \,\}$$

Read that as a picture

The confidence interval is the **horizontal slice** of the p-value curve at height α.

The p-Value Curve, Live

Drag θ_0 and watch the test on the left agree with the interval on the right — every time.

θ_0

0.4

α

0.05

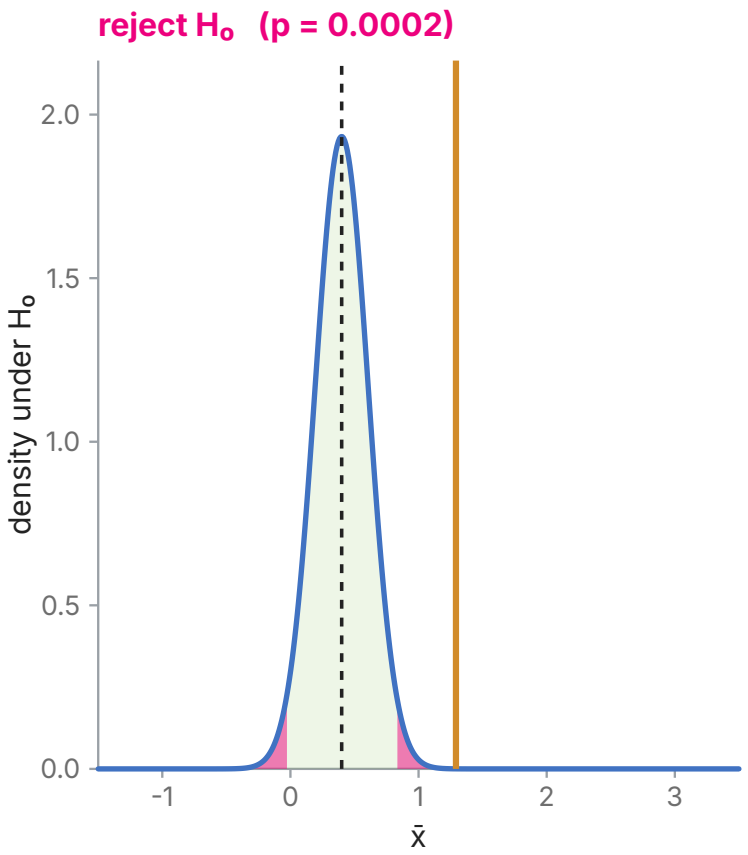
n

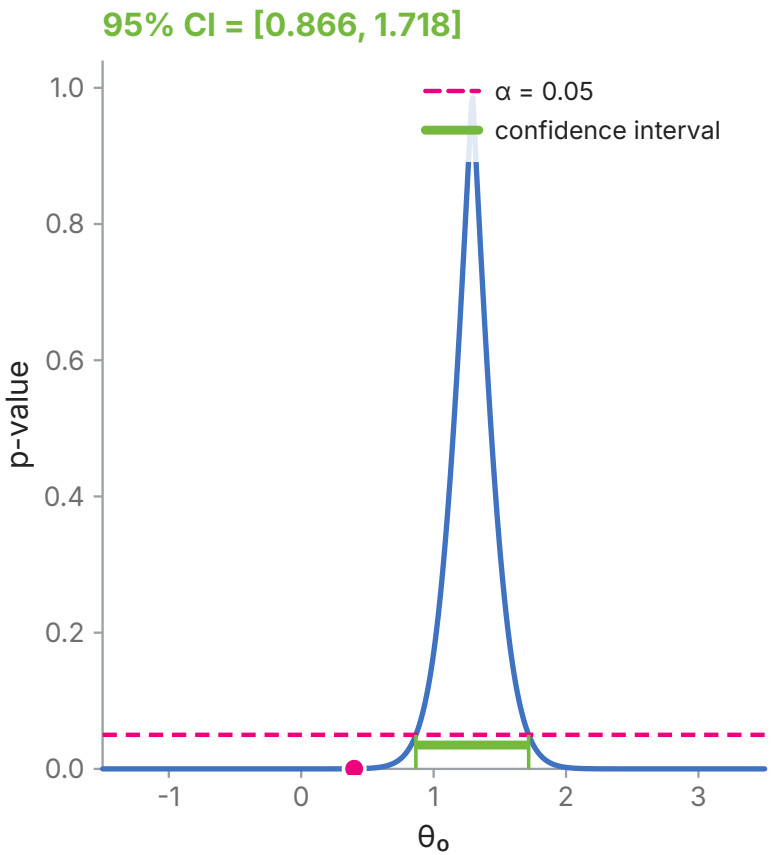
25

New sample

Left: the null distribution centred at the current θ_0, with the observed $\bar{x}$ and its tail area.

Right: $p(\theta_0)$ for **every** θ_0. Where the curve crosses α are exactly the interval endpoints.





What the Duality Buys You

- You never need two sets of tables. “Is θ_0 in the interval?” and “is $p < \alpha$?” are one question
- Lower α and the slice moves down, so the interval **widens** — the same trade-off a rejection region draws
- The interval says strictly more than the test: it also reports every value the data **cannot** rule out

Why this matters for reporting

“Not significant” collapses an interval to one bit. An interval of $[-0.01,\ 0.42]$ and an interval of $[-0.20,\ 0.21]$ tell very different stories, and the p-value is nearly the same for both.

When the Formula Runs Out

The Luxury You Only Have in a Lecture

Every simulation so far drew fresh samples **from a known truth** — a luxury you have exactly once, in a lecture.

You cannot do that with real data

One sample, no population to redraw from. The way out so far was a **pivot** with a known distribution; these have none:

- The **Gini coefficient**, the 90–10 ratio, top income shares
- **Elasticities**, ratios of estimates, differences of **quantiles**
- Anything computed by a multi-step procedure

The bootstrap principle

The sample is to the population as a **resample** is to the sample. Draw B samples of size n **with replacement** from the data, recompute $\hat{\theta}^*$ on each, and use the spread of $\hat{\theta}^*$ as the sampling distribution of $\hat{\theta}$.

The Bootstrap, Live

Statistic

Gini coefficient

Sample size n

80

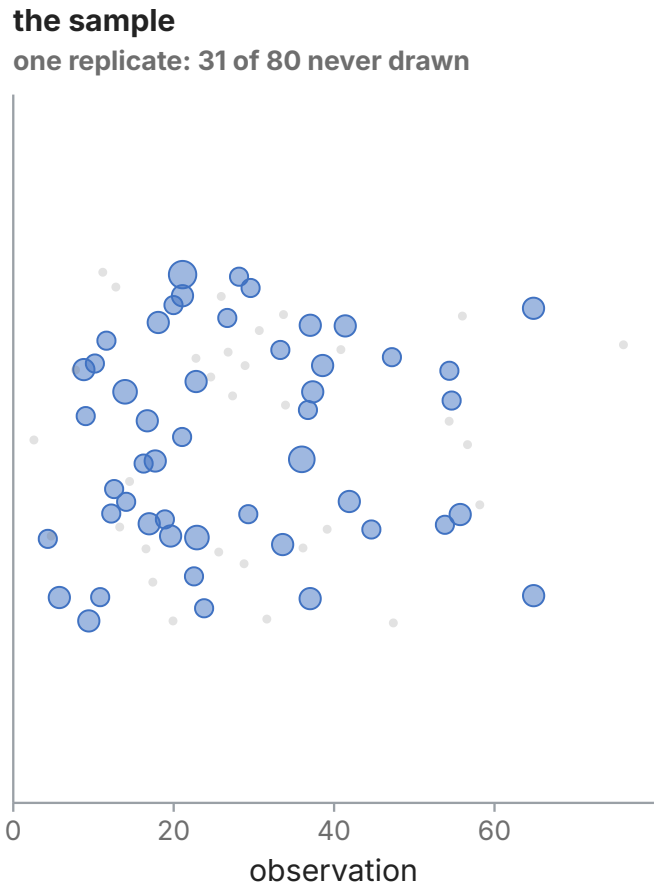
Replicates B

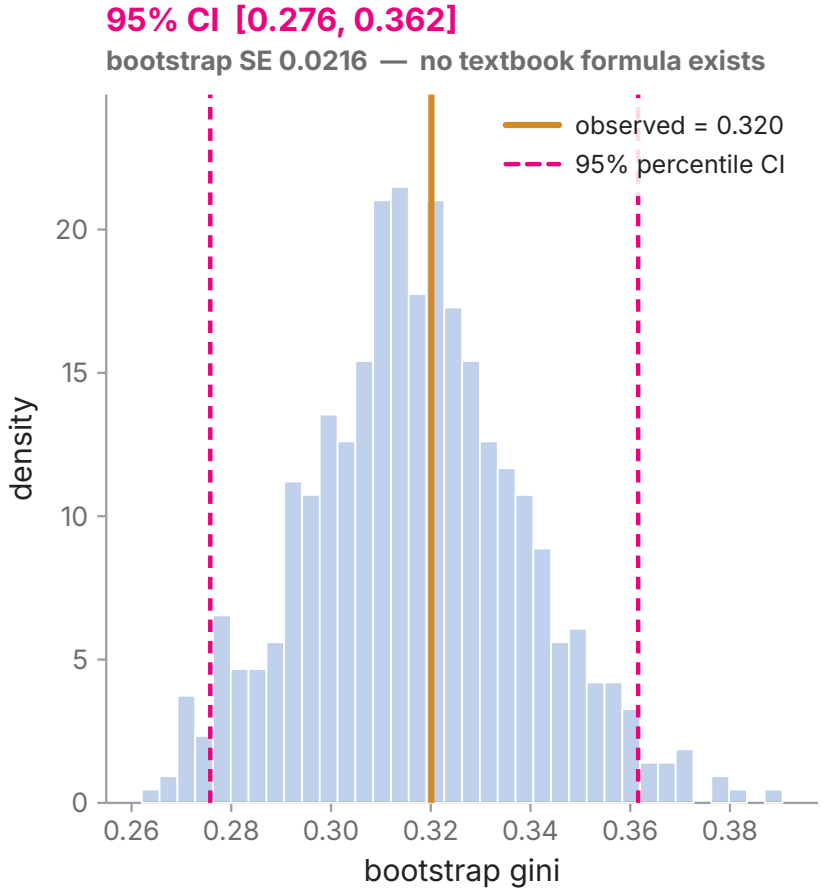
600

New sample

Left: the sample. Dot **size** is how often each observation was drawn in one replicate; greyed-out points were not drawn at all.

Choose **Maximum** to watch the method fail: a resample misses the largest observation with probability $(1 - 1/n)^n \to e^{-1}$.





What the Bootstrap Can and Cannot Do

- For the **mean**, the bootstrap standard error reproduces $s/\sqrt{n}$ to two decimals — which is what earns it trust
- For the **Gini**, it produces an interval where no textbook formula exists at all
- For the **maximum** it collapses: a resample misses the largest observation with probability $(1 - 1/n)^n \to e^{-1} \approx 0.368$, so roughly 63% of replicates return exactly the sample max

B is not n

Raising B shrinks Monte-Carlo noise in the **endpoints**. It does not shrink the interval, and it adds no information.

It cannot repair the sample

Resampling a biased sample gives biased resamples. And i.i.d. resampling destroys dependence — clustered or serial data need a block bootstrap.

Back to the Estimator: Attaining the Bound

Picking Up the Other Road

The CRLB said a floor exists. It did not say anything reaches it. The rest of this deck is the machinery that does — and one method for when even the likelihood is more than you are willing to assume.

Sufficiency → Rao-Blackwell → UMVUE

A recipe that provably cannot be beaten among unbiased estimators.

GMM

What to do when you can write down moments but not a density.

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

n (flips)

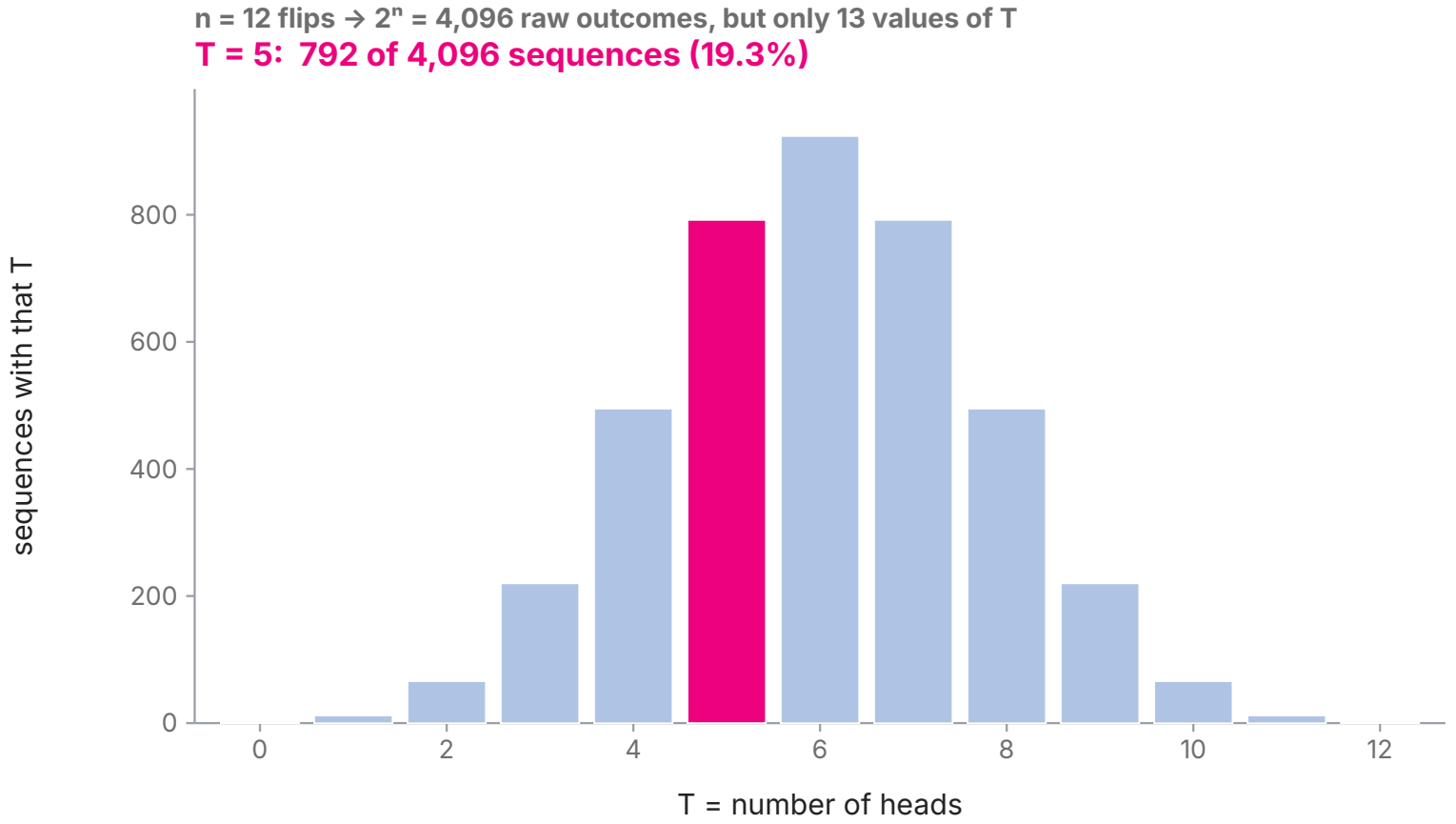
12

T (highlighted)

5

All sequences with the same count $T = \sum x_i$ are **interchangeable** for learning p. The bar is how many raw sequences collapse into that one value.

2^n possible datasets, $n + 1$ possible values of T — and nothing about p is lost in the collapse.



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

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

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

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: $\hat{p}^* = \mathbb{E}[X_1 \mid T] = T/n$

Result

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

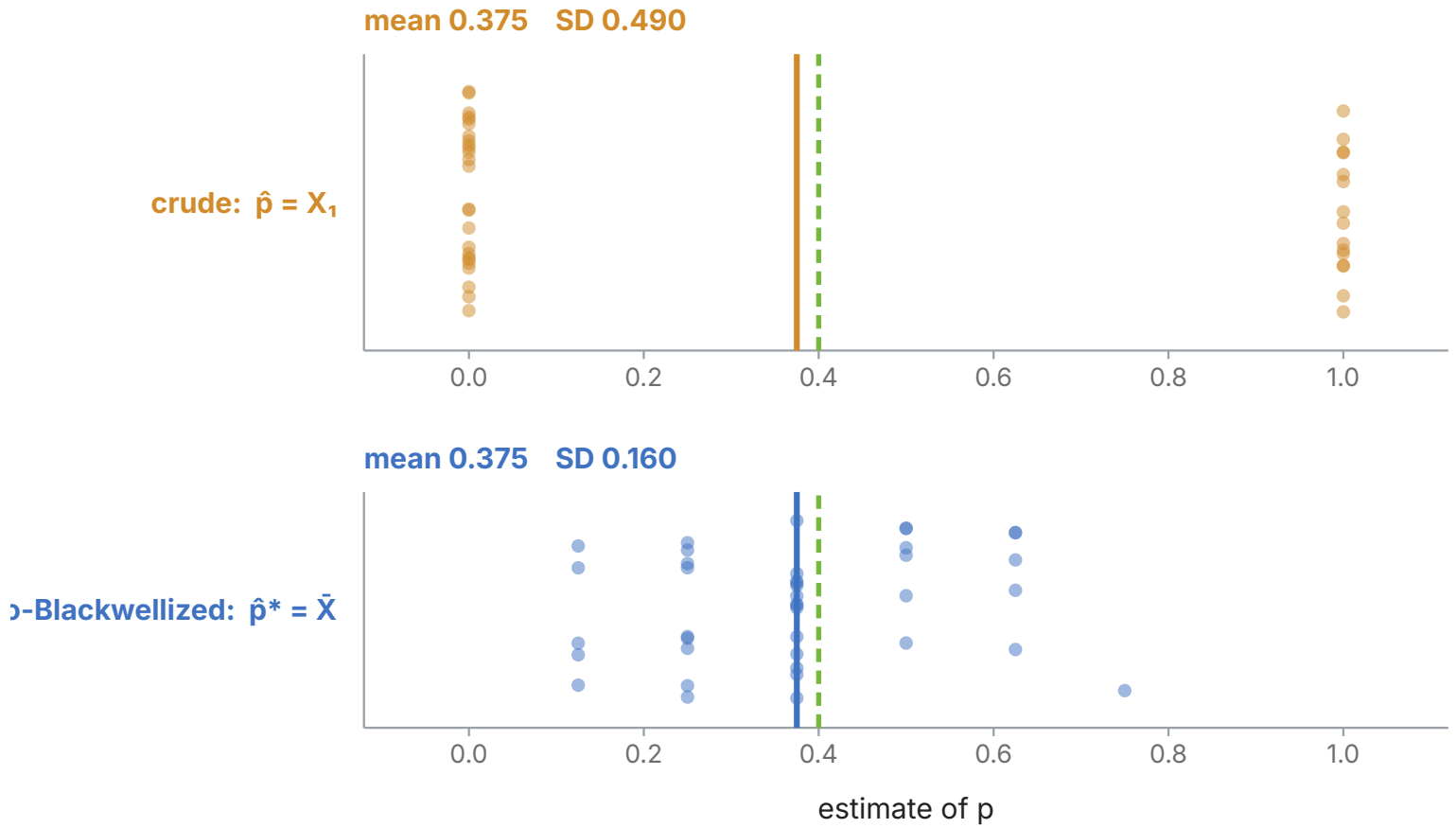
Rao–Blackwell, Visualized

repeated experiments

40

Each dot is one experiment of $n = 8$ coin flips with true $p = 0.4$.

Both rows are centred on p — both estimators are unbiased. Only the **spread** differs, and that spread is the standard error. Theory says it should fall by $\sqrt{n} = \sqrt{8} \approx 2.83$: conditioning on T discarded noise, not information.



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

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

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

And it need not sit on the floor

The CRLB is a bound the UMVUE may or may not attain — and the shrinkage slide showed a **biased** estimator beating it on MSE. "Best unbiased" is not "best".

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:

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

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 idea as the Cramér-Rao bound, transplanted from the likelihood to the moment conditions.

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 $\mathbf{E}[X'\epsilon] \neq 0$ — X is **endogenous**, so OLS is biased.

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

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

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.

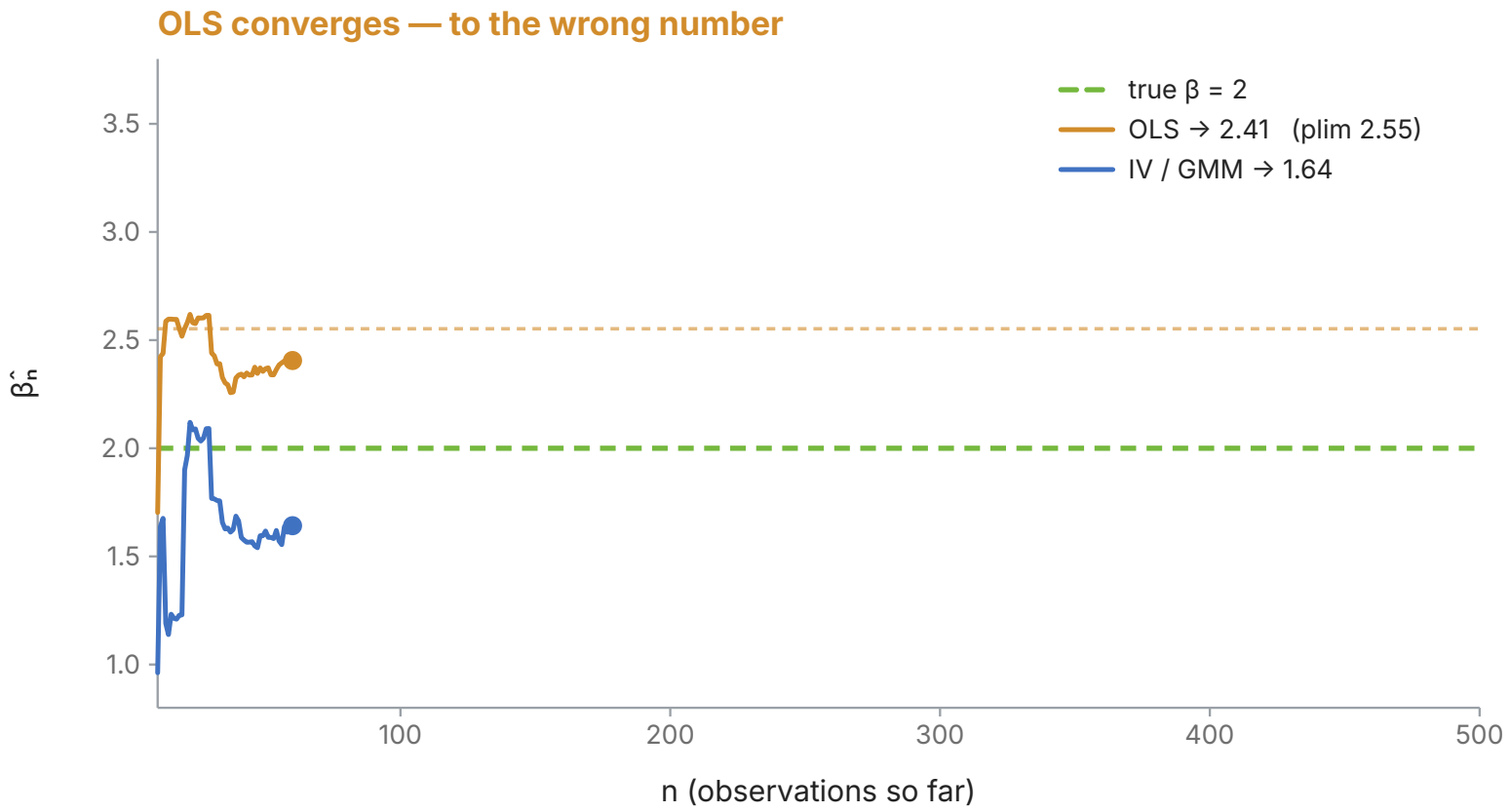
OLS Bias vs. IV Consistency, Visualized

sample size n

60

A live simulation with $x = 0.9z + u$ and $y = 2x + u + \varepsilon$: the confounder u sits in both, so x is endogenous.

Both lines are noisy at small n and both settle down — but they settle on different numbers. More data does not cure a biased estimator; it only pins down the wrong answer more precisely.



GMM Shows Up All Over Economics

- **Instrumental variables & 2SLS:** fixing endogeneity in regression (the last two slides)
- **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

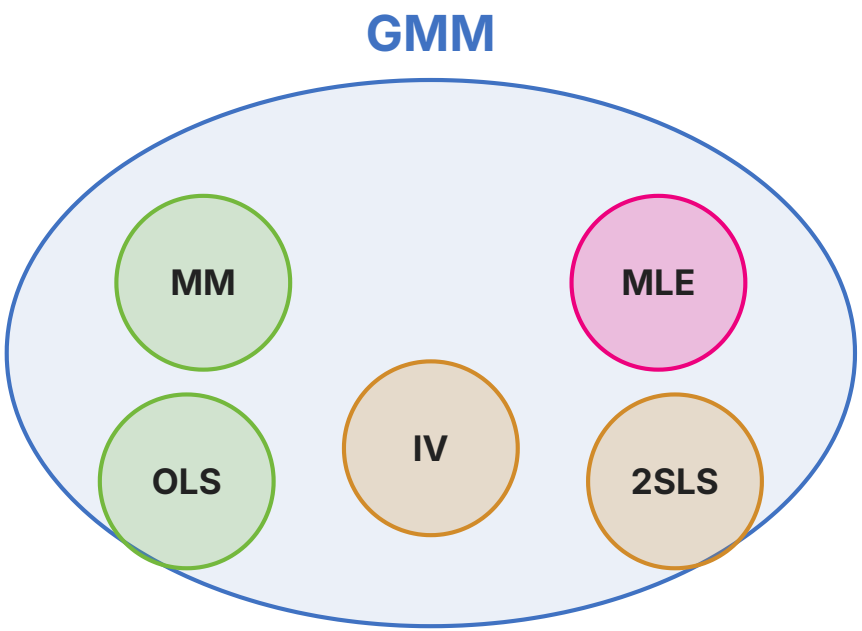
J-Test for Over-Identification

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.

And a test is an interval seen from the other side: the values of θ the J-test does not reject are a confidence region.



each is GMM with particular moments 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

Summary

Comparing the Estimation Methods

Property	GMM	LS	MLE	UMVUE
Always exists	✅	✅	✅	❌
Unbiased	❌	✅[1]	❌	✅
Consistent	✅	✅	✅	✅
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 — The Estimator

(1) An estimator is judged by its **sampling distribution**: where it sits (bias) and how wide it is (variance)

(2) **MSE = variance + bias²** — which is why a little bias can be worth buying

(3) **Consistency** is a promise about $n \to \infty$; unbiasedness is a promise at fixed n. Neither implies the other

(4) The **Cramér-Rao bound** is a floor on the variance of any unbiased estimator

(5) **Sufficiency**, **Rao-Blackwell** and **completeness** are the recipe for reaching it; **GMM** is what to use when you cannot write down a density

The through-line

Every one of these is a statement about the **same** distribution — the one you never observe, because you only ever draw one sample from it.

Key Takeaways — The Interval

(1) A confidence interval is a statement about the **procedure**; coverage is kept only while its assumptions hold

(2) **Pivots** generate intervals — and only symmetric pivots give $\hat{\theta} \pm c \cdot \mathrm{SE}$

(3) An interval **is** an inverted test: the slice of the p-value curve at height α

(4) Precision costs $n \propto w^{-2}$ — the Cramér-Rao bound, priced in observations

(5) The **bootstrap** covers statistics with no formula — and fails visibly at the edge of the sample

Choosing an interval

- **Mean, unknown σ** — t interval
- **Proportion** — Wilson, never Wald
- **Variance** — χ^2, asymmetric
- **Smooth function of estimates** — delta method
- **No formula** — bootstrap

Report the interval

It carries the effect size, the precision, and every value the data cannot rule out. A p-value carries one bit of that.

What This Deck Left Out

Bayesian methods

Priors and posteriors · **credible** intervals, the probability statement a CI is not · hierarchical models

Robust estimation

M-, L- and R-estimators · influence functions

Modern extensions

High-dimensional estimation · shrinkage with LASSO and Ridge · machine learning and causal inference

And the interval kind

The **delta method** for functions of estimates · simultaneous and uniform bands · block and cluster bootstraps