Lesson 3 — Estimator Properties

Consistency, Bias, Variance, Confidence Intervals

Applied Statistics Course

Outline

Bias–Variance Tradeoff

Bias–Variance Tradeoff

Learning Objectives

- Define bias, variance, and mean squared error of estimators
- Explain the bias–variance decomposition: $\text{MSE} = \text{Bias}^2 + \text{Var}$
- Apply tradeoff concepts to shrinkage estimators and sample variance
- Connect to Lesson 2 estimator properties and Lesson 1 sampling distributions

Builds on Lesson 2 (MLE/MoM) foundations

Bias

Definition

The bias of an estimator is

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

Intuition

Bias measures systematic error: if non-zero the estimator is centered away from the true parameter even as we average over repeated samples. In practice bias affects accuracy and may require correction or a bias–variance tradeoff.

Variance

Definition

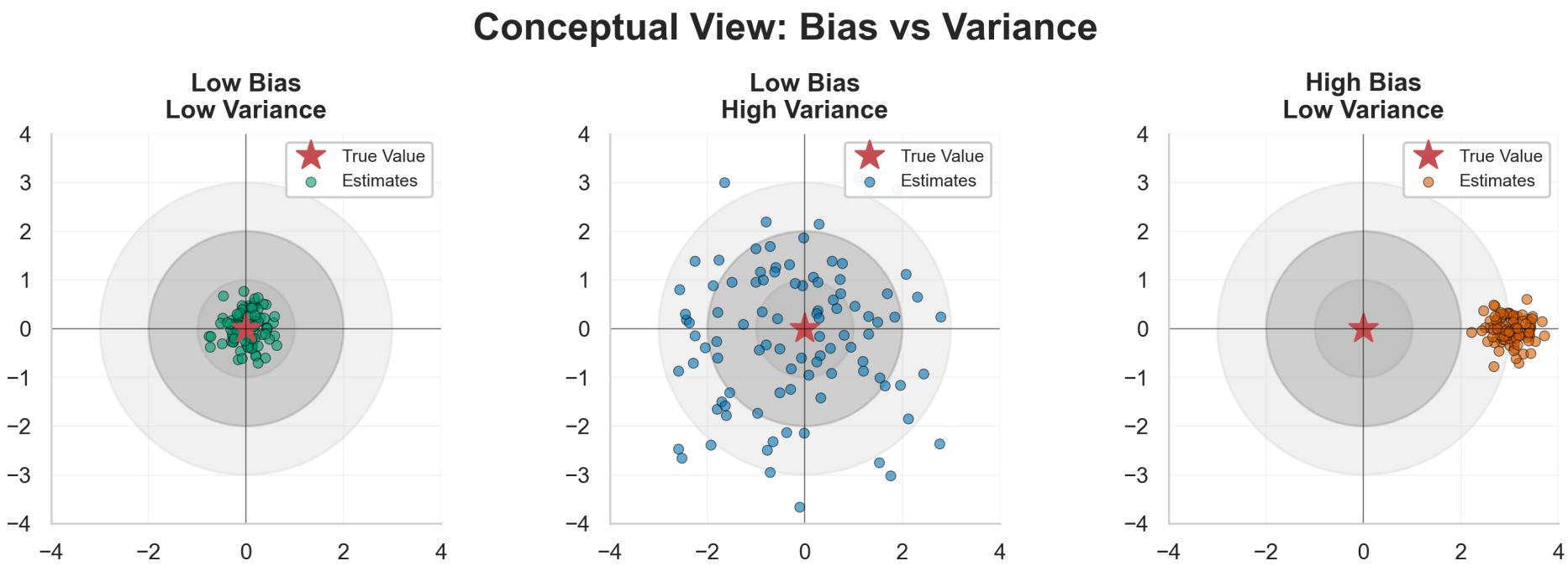
The variance of an estimator is

$$\mathrm{Var}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \mathbb{E}[\hat{\theta}])^2\big].$$

Intuition

Variance measures randomness in the estimator across different samples. Low variance means repeated experiments produce similar estimates; high variance implies lack of precision. For practitioners, variance controls confidence interval width and sample size planning.

Variance — Visual



MSE and Bias–Variance Tradeoff

Definition

The mean squared error of an estimator is

$$\mathrm{MSE}(\hat{\theta}) = \mathbb{E}\big[(\hat{\theta} - \theta)^2\big].$$

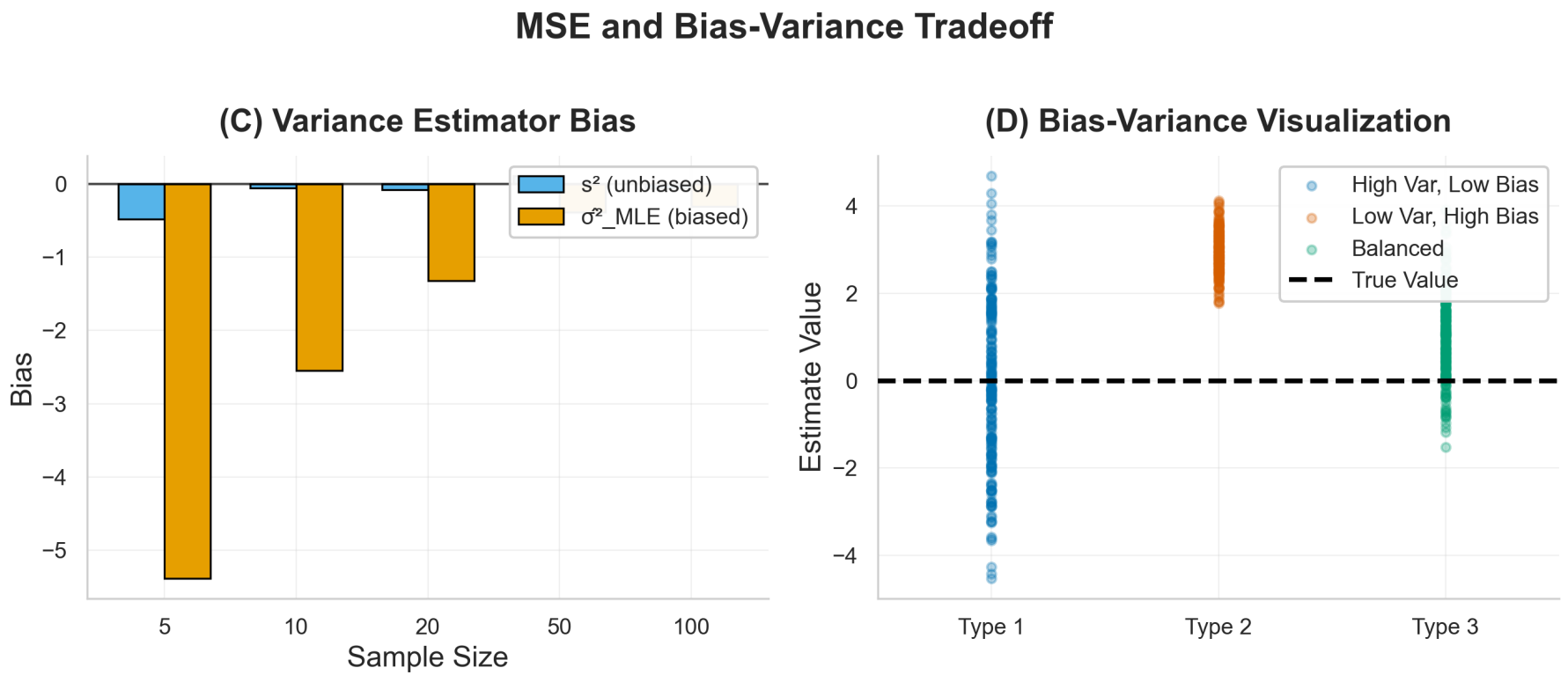
It decomposes as

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{Bias}(\hat{\theta})^2 + \mathrm{Var}(\hat{\theta}).$$

Intuition

MSE trades off accuracy (bias) against precision (variance). In applied work a small bias can be acceptable if it substantially reduces variance and thereby improves prediction or interval width.

MSE and Bias–Variance Tradeoff — Visual



Left: Variance estimator bias comparison; Right: Visual representation of bias-variance tradeoff scenarios

Shrinkage Estimator

Definition

The shrinkage estimator combines data and prior information:

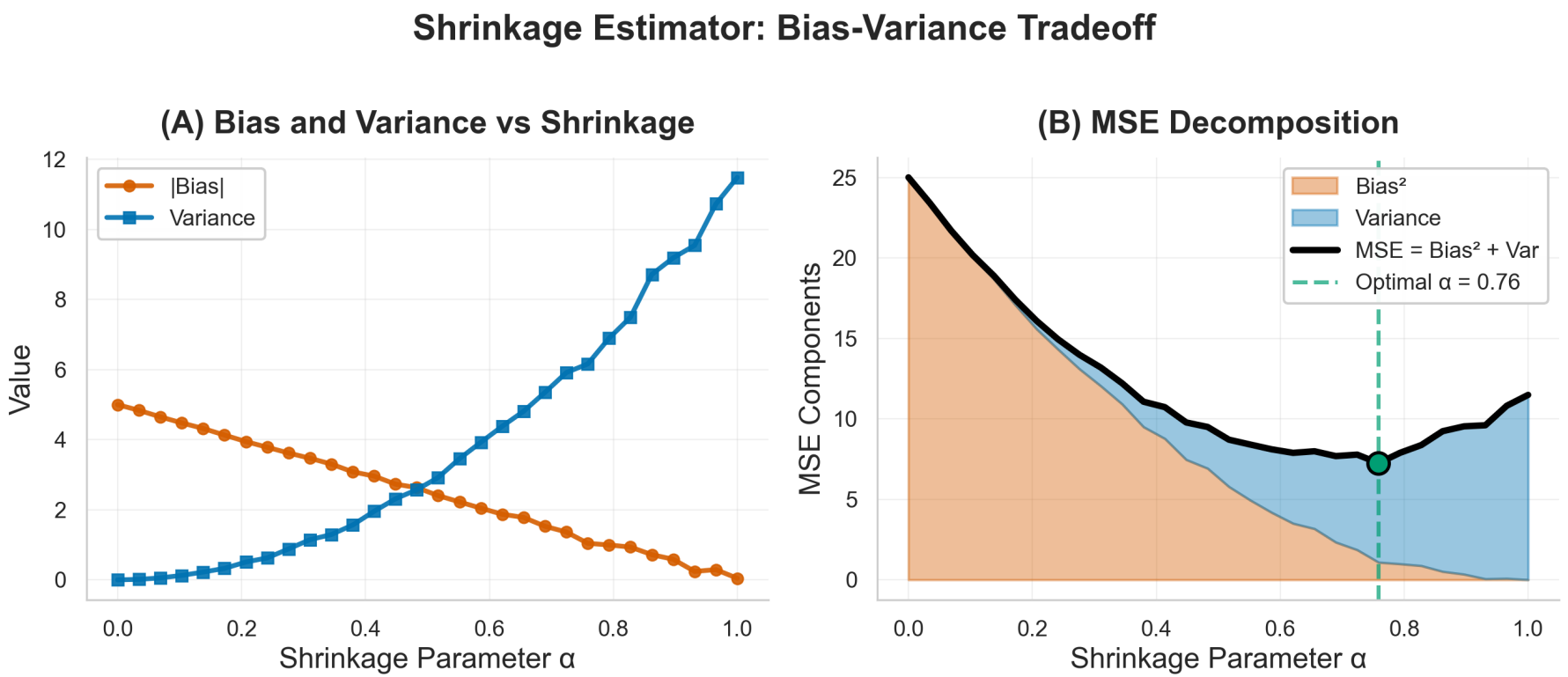
$$\delta_\alpha = \alpha \bar{X} + (1 - \alpha)\mu_0,$$

where μ_0 is a prior guess and $\alpha \in [0, 1]$ controls shrinkage.

Connection to Lesson 2

Builds on MLE/MoM estimators by showing how to incorporate prior information when data is limited, using heights data from `shared/data/heights_weights_sample.csv`.

Shrinkage Estimator — Visual



Left: Bias and variance as functions of α; Right: MSE decomposition showing optimal shrinkage

Sample Variance Estimators

Definition

Two common variance estimators (building on Lesson 2 Normal examples):

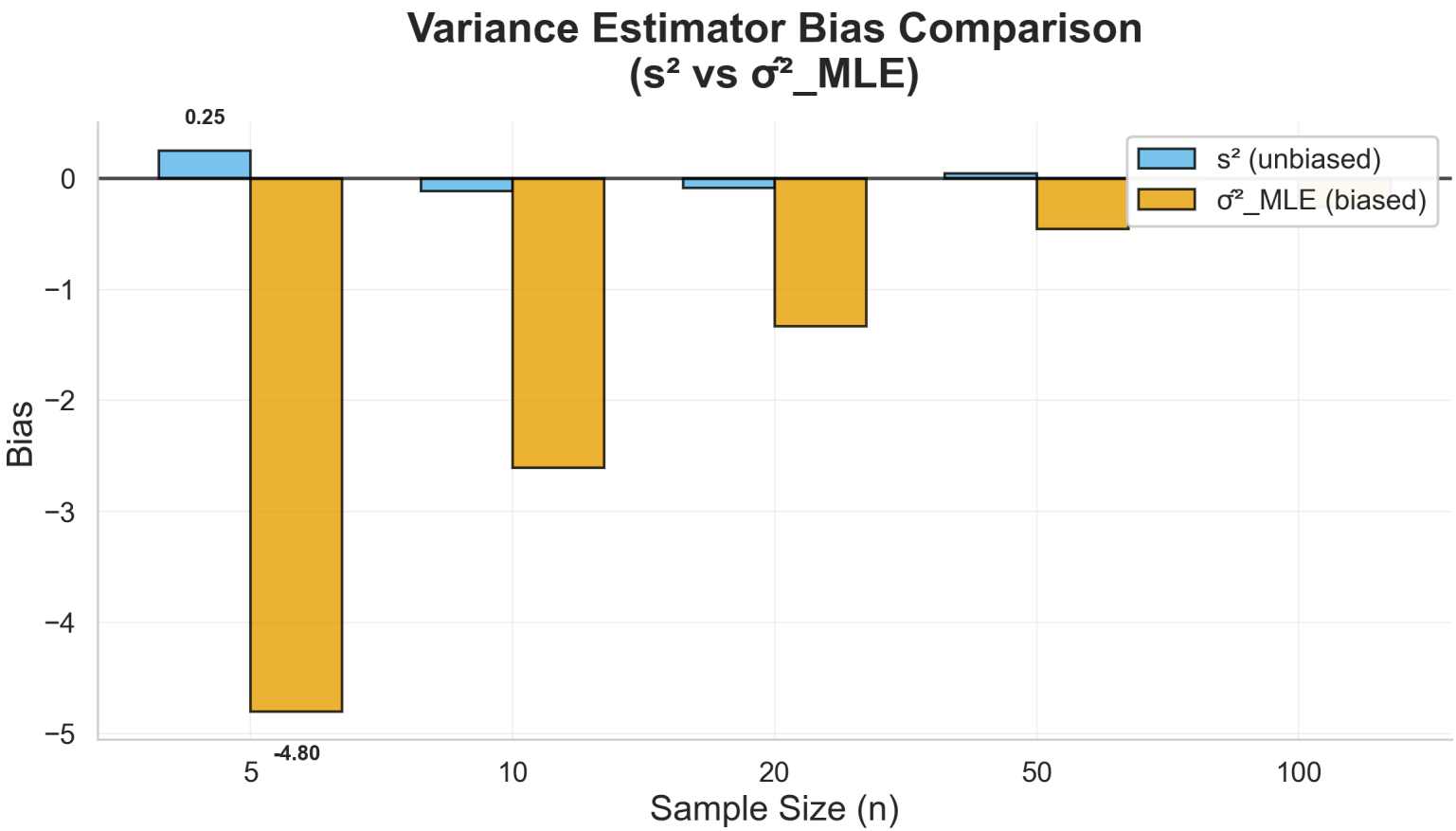
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{(unbiased)},$$

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{(MLE, biased)}.$$

Tradeoff Analysis

The unbiased estimator has lower bias but higher variance; the MLE has higher bias but lower variance, illustrating the bias–variance tradeoff.

Sample Variance Estimators — Visual



Comparison shows s^2 is unbiased while $\hat{\sigma}^2_{\text{MLE}}$ has negative bias that decreases with n

Pitfalls & Heuristics

Common Pitfalls

- Overlooking bias in small samples
- Assuming unbiased = better (ignores variance)
- Not considering the use case (prediction vs estimation)

Practical Heuristics

- Use unbiased estimators when bias matters most
- Consider shrinkage when prior information is reliable
- Evaluate estimators on MSE for prediction tasks
- Check both bias and variance in simulation studies

Exercises

1. Derive the bias of the MLE variance estimator $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{n}\sum(X_i - \bar{X})^2$ for Normal data.
2. Show that $\text{MSE}(\hat{\theta}) = \text{Bias}(\hat{\theta})^2 + \text{Var}(\hat{\theta})$ using the definition of variance.
3. Simulate the bias–variance tradeoff for the shrinkage estimator $\delta_\alpha = \alpha\bar{X} + (1-\alpha)\mu_0$ with $\mu_0 = 170$ using heights data.
4. Compare MSE of s^2 vs $\hat{\sigma}^2_{\text{MLE}}$ across different sample sizes $n = 5, 10, 20, 50$.

Summary

Key Takeaways

- Bias measures systematic error; variance measures random error
- $\text{MSE} = \text{Bias}^2 + \text{Var}$ shows the fundamental tradeoff
- Unbiased estimators aren't always better (consider variance)
- Shrinkage can reduce variance at the cost of some bias
- Choice depends on context: estimation vs prediction

Connects Lesson 2 estimators to practical performance evaluation

- Wikipedia: Bias–variance tradeoff
  `https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff`
- Casella & Berger, *Statistical Inference* (Chapter 7)

Outline

Consistency

Consistency

Learning Objectives

- Define consistency (convergence in probability) and strong consistency
- Connect consistency to the Law of Large Numbers (Lesson 1)
- Identify consistent vs inconsistent estimators
- Apply consistency concepts to MLE and MoM estimators (Lesson 2)

Builds on Lesson 1 (LLN/CLT) and Lesson 2 (MLE/MoM)

Consistency

Definition

An estimator $\hat{\theta}_n$ is consistent for θ if

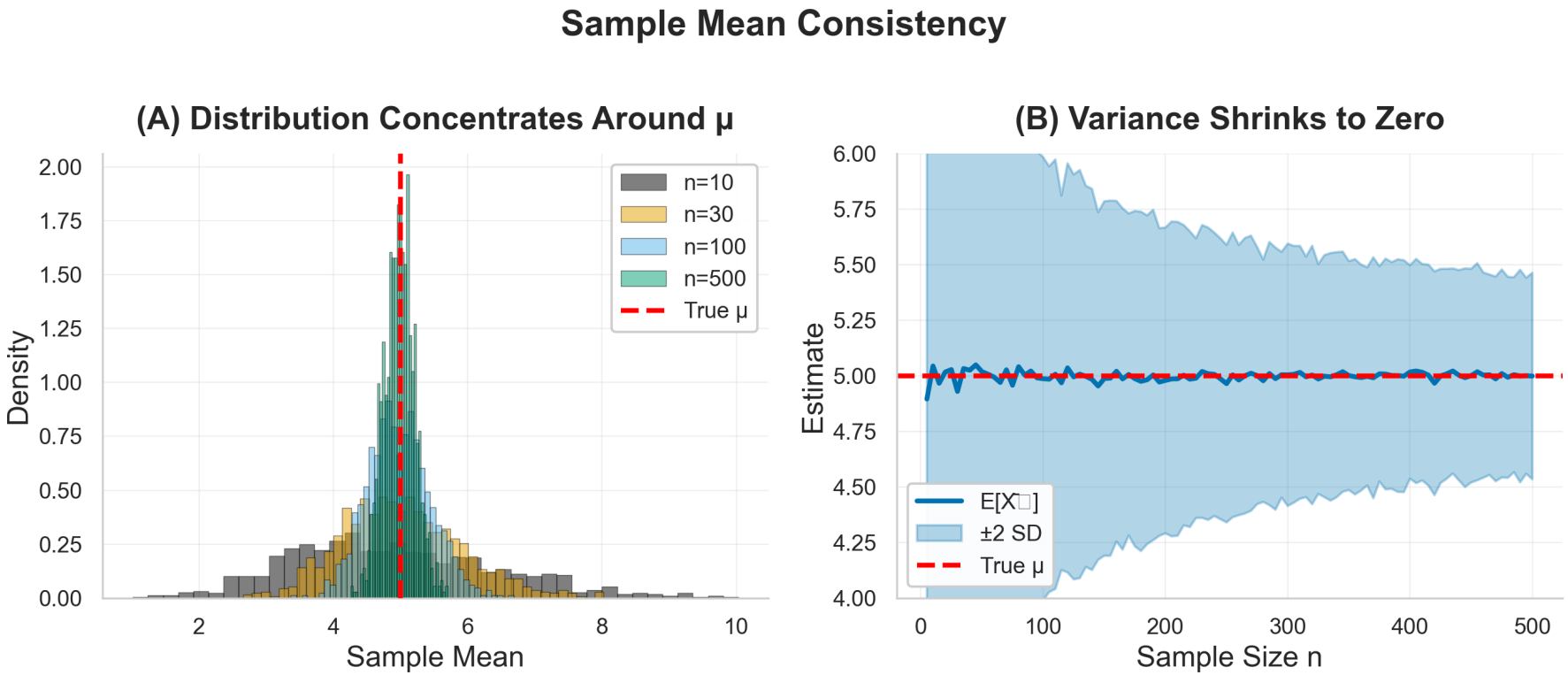
$$\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta \qquad (n \to \infty),$$

meaning $\forall \epsilon > 0$, $\mathbb{P}(|\hat{\theta}_n - \theta| > \epsilon) \to 0$.

Intuition

With increasing sample size the estimator concentrates around the true value. Consistency is a minimal long-run requirement: without it an estimator may never learn the truth no matter how much data you collect.

Consistency — Visual



Left: Sample mean distributions concentrate around as n increases; Right: Variance shrinks to zero

Strong Consistency

Definition

An estimator $\hat{\theta}_n$ is strongly consistent if

$$\hat{\theta}_n \xrightarrow{\text{a.s.}} \theta \qquad (n \to \infty),$$

meaning $\mathbb{P}(\lim_{n\to\infty} \hat{\theta}_n = \theta) = 1$.

Connection to Lesson 1

Strong consistency follows from the Strong Law of Large Numbers, while weak consistency follows from the Weak Law of Large Numbers.

Law of Large Numbers and Consistency

Weak LLN $\rightarrow$ Consistency

For i.i.d. data with $\mathbb{E}[X_i] = \mu < \infty$:

$$\bar{X}_n \xrightarrow{\text{P}} \mu \quad \Rightarrow \quad \bar{X}_n \text{ is consistent for } \mu.$$

Strong LLN $\rightarrow$ Strong Consistency

Under the same conditions:

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu \quad \Rightarrow \quad \bar{X}_n \text{ is strongly consistent for } \mu.$$

Practical Implication

Sample means are consistent estimators of population means, justifying the use of $\bar{X}_n$ as an estimator for μ in large samples.

Example: Sample Mean (Consistent)

Normal Case

For $X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d.:

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{P}} \mu.$$

Connection to Lesson 1

This follows directly from the Weak Law of Large Numbers and the Central Limit Theorem, providing the foundation for statistical inference.

Why It Works

- $\mathbb{E}[\bar{X}_n] = \mu$ (unbiased)
- $\text{Var}(\bar{X}_n) = \sigma^2/n \to 0$ (variance vanishes)
- By Chebyshev: $\mathbb{P}(|\bar{X}_n - \mu| > \epsilon) \leq \text{Var}(\bar{X}_n)/\epsilon^2 \to 0$

Example: Uniform Maximum (Biased but Consistent)

Setup

Let $X_i \sim \mathcal{U}[0, \theta]$ i.i.d., and consider

$$\hat{\theta}_n = \max\{X_1, \ldots, X_n\}.$$

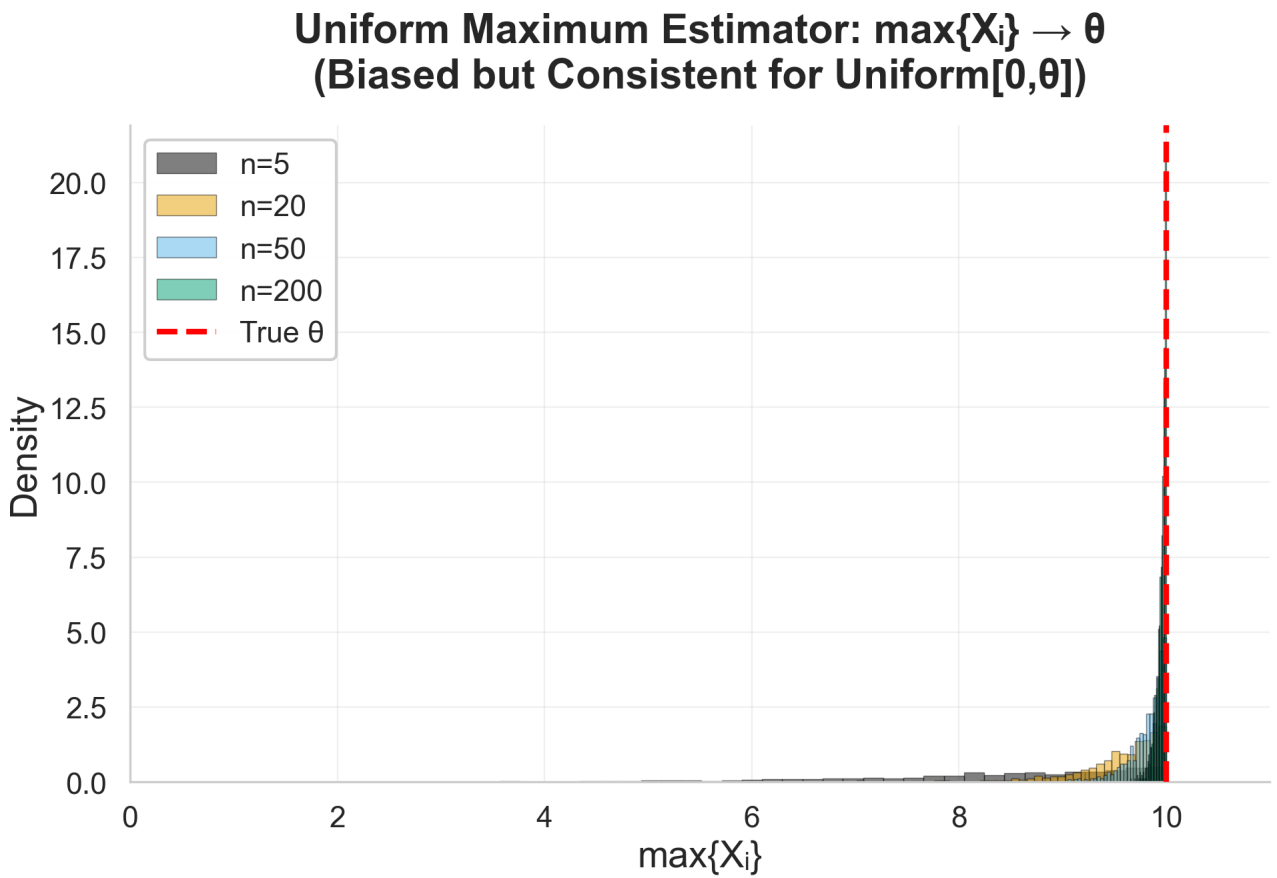
Properties

- $\mathbb{E}[\hat{\theta}_n] = \frac{n}{n+1}\theta < \theta$ (biased)
- $\mathrm{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$
- $\hat{\theta}_n \xrightarrow{\mathrm{P}} \theta$ (consistent)

Connection to Lesson 2

This estimator is the MLE for the Uniform$[0, \theta]$ distribution, demonstrating that biased estimators can still be consistent.

Uniform Maximum — Visual



Distribution of max{X} concentrates at as n increases, demonstrating consistency despite bias

Example: Inconsistent Estimator

Counterexample

Consider using X_1 (the first observation) as an estimator for μ:

$$\hat{\mu}_n = X_1.$$

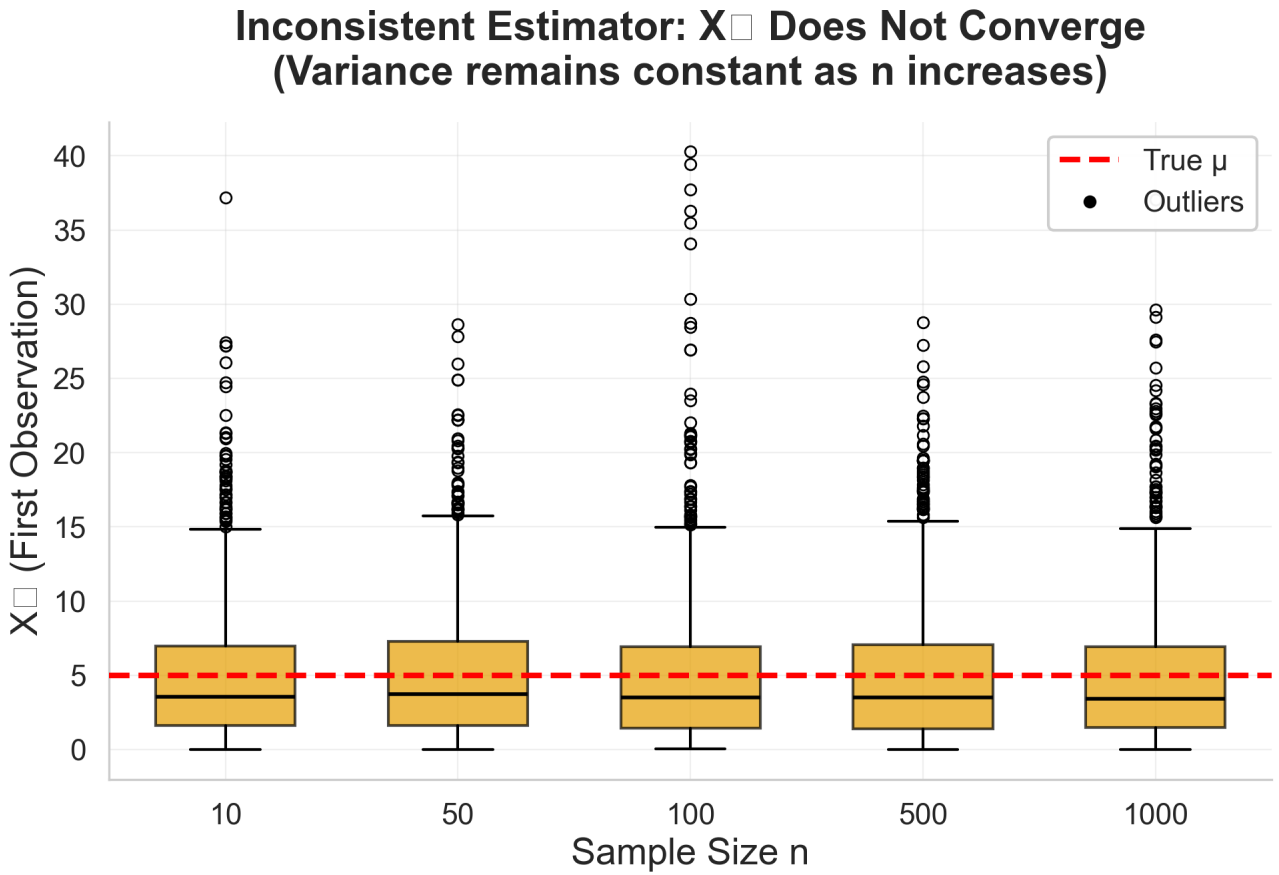
Why Inconsistent

- $\mathbb{E}[\hat{\mu}_n] = \mu$ (unbiased)
- $\text{Var}(\hat{\mu}_n) = \sigma^2$ (variance doesn't decrease with n)
- $\mathbb{P}(|\hat{\mu}_n - \mu| > \epsilon) = \mathbb{P}(|X_1 - \mu| > \epsilon) \not\to 0$

Contrast with Lesson 1

This violates the variance vanishing requirement for consistency, unlike the sample mean which satisfies $\text{Var}(\bar{X}_n) \to 0$.

Inconsistent Estimator — Visual



Distribution of X remains unchanged regardless of sample size n

Pitfalls: Heavy-Tailed Distributions

LLN Failure

The Law of Large Numbers requires finite variance. For heavy-tailed distributions with $\text{Var}(X_i) = \infty$, the sample mean may not be consistent.

Example

For $X_i \sim$ Cauchy distribution:

- No finite mean or variance
- Sample mean does not converge (in probability or a.s.)
- Need robust alternatives (median, trimmed mean)

Practical Implication

Always check moment conditions before assuming consistency, especially with real data that may have heavy tails.

Exercises

1. Prove that if $\hat{\theta}_n \xrightarrow{\text{P}} \theta$ and g is continuous, then $g(\hat{\theta}_n) \xrightarrow{\text{P}} g(\theta)$.
2. Show that the sample median is consistent for the population median under mild conditions.
3. Use the `uniform_max_consistency()` function from the appendix to verify that $\max\{X_i\}$ is consistent for θ in Uniform$[0, \theta]$.
4. Explain why X_1 is inconsistent for μ while $\bar{X}_n$ is consistent.

Summary

Key Takeaways

- Consistency requires both correct centering and vanishing variance
- Sample means are consistent by LLN (Lesson 1 foundation)
- Biased estimators can be consistent if bias vanishes
- MLEs and MoM estimators are typically consistent (Lesson 2)
- Heavy tails can break consistency assumptions

Provides theoretical foundation for Lesson 2 estimators

- Wikipedia: Consistent estimator
  `https://en.wikipedia.org/wiki/Consistent_estimator`
- Casella & Berger, *Statistical Inference* (Chapter 7)

Asymptotic Efficiency & CRLB

Asymptotic Efficiency & Cramér–Rao Lower Bound

Learning Objectives

- Define score function, Fisher information, and CRLB for unbiased estimators
- State regularity conditions and equality cases
- Explain asymptotic normality of MLEs and asymptotic efficiency
- Work through Normal, Poisson, and Exponential examples (Lesson 2)

Extends Lesson 2 (Fisher information, MLE) foundations

Score Function

Definition

The score function is the derivative of the log-likelihood:

$$U(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}.$$

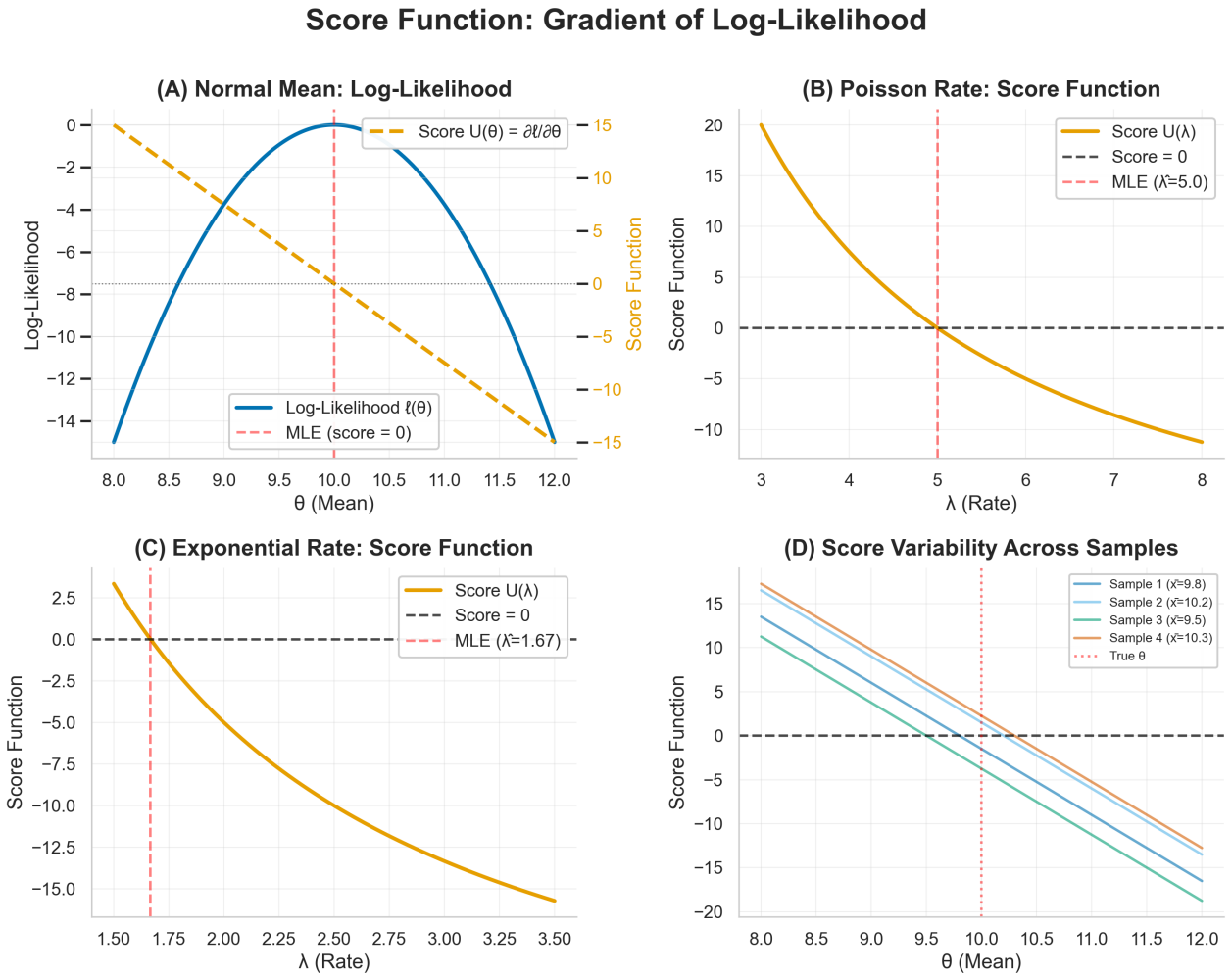
Intuition

The score measures the sensitivity of the log-likelihood to changes in the parameter. It indicates the direction and magnitude of the gradient that the MLE will follow.

Connection to Lesson 2

For i.i.d. data, the total score is $U_n(\theta) = \sum_{i=1}^{n} U_i(\theta)$, and the MLE satisfies $U_n(\hat{\theta}_{\text{MLE}}) = 0$.

Score Function — Visual



Takeaway: The score function is zero at the MLE, indicating the likelihood peak.

Fisher Information

Definition

The Fisher information is the variance of the score:

$$\mathcal{I}(\theta) = \mathrm{Var}(U(\theta)) = -\mathbb{E}\left[\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right].$$

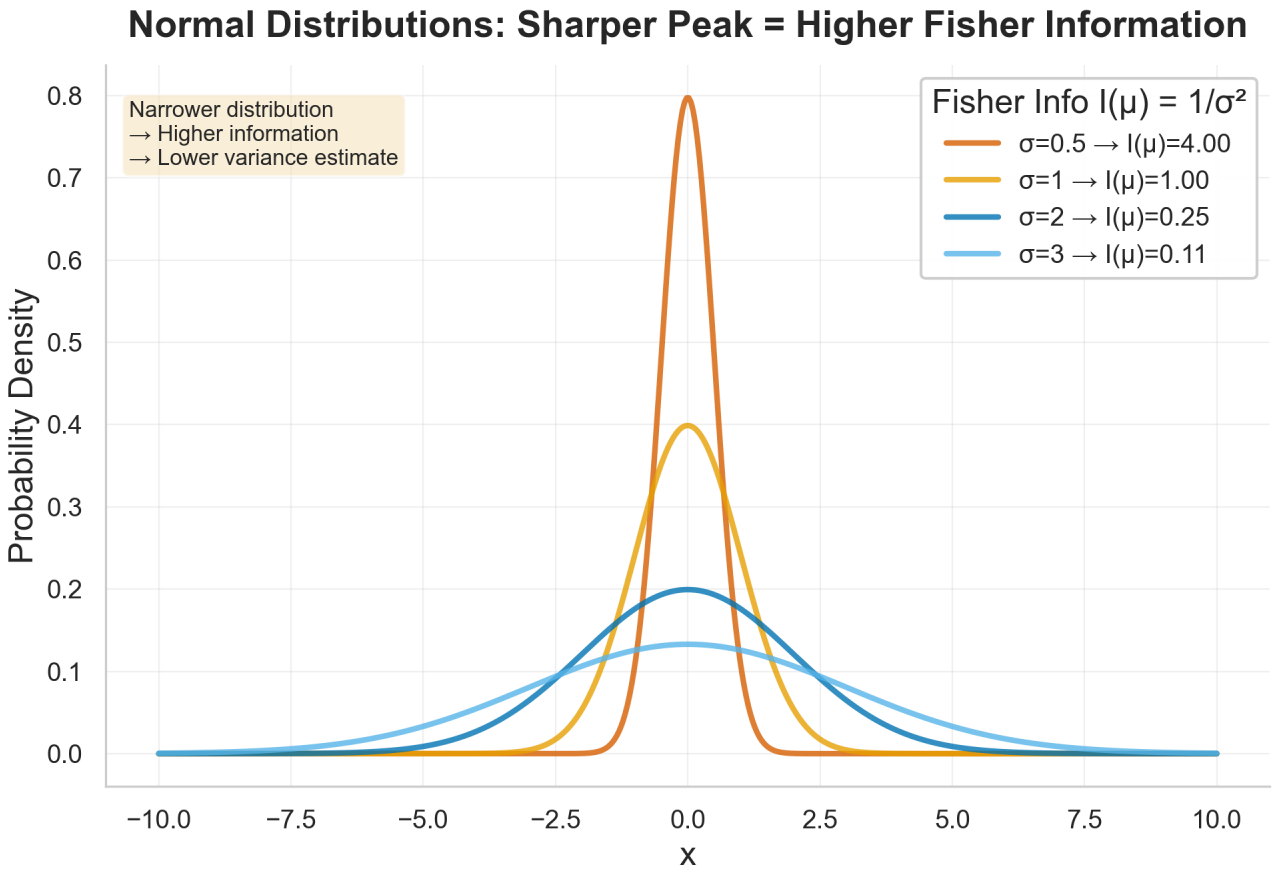
Intuition

Fisher information quantifies the amount of information about θ contained in the data. Higher information means sharper likelihood peaks and lower achievable variance.

Connection to Lesson 2

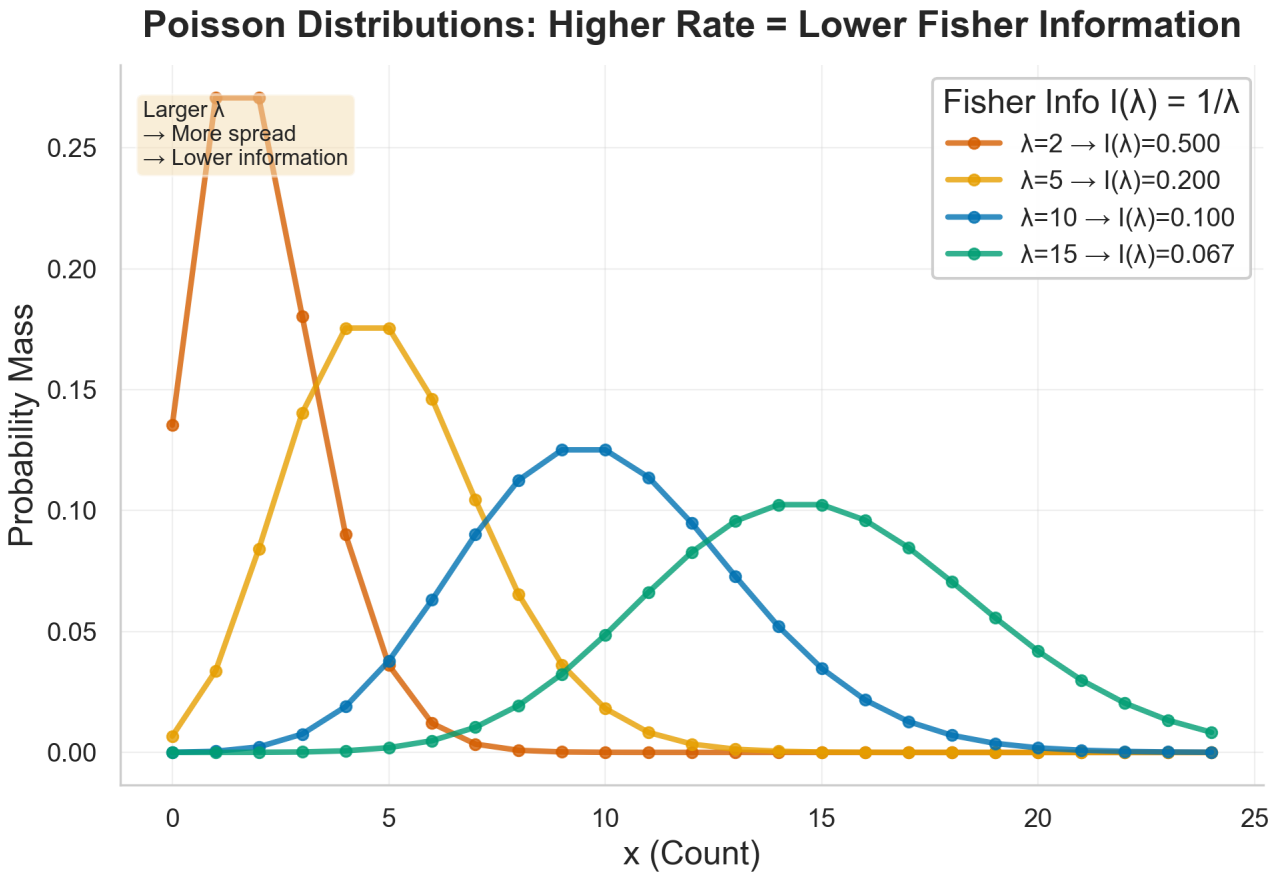
For i.i.d. data, the total information is $\mathcal{I}_n(\theta) = n\mathcal{I}(\theta)$, explaining the $1/n$ scaling of MLE variance.

Fisher Information — Visual (Normal)



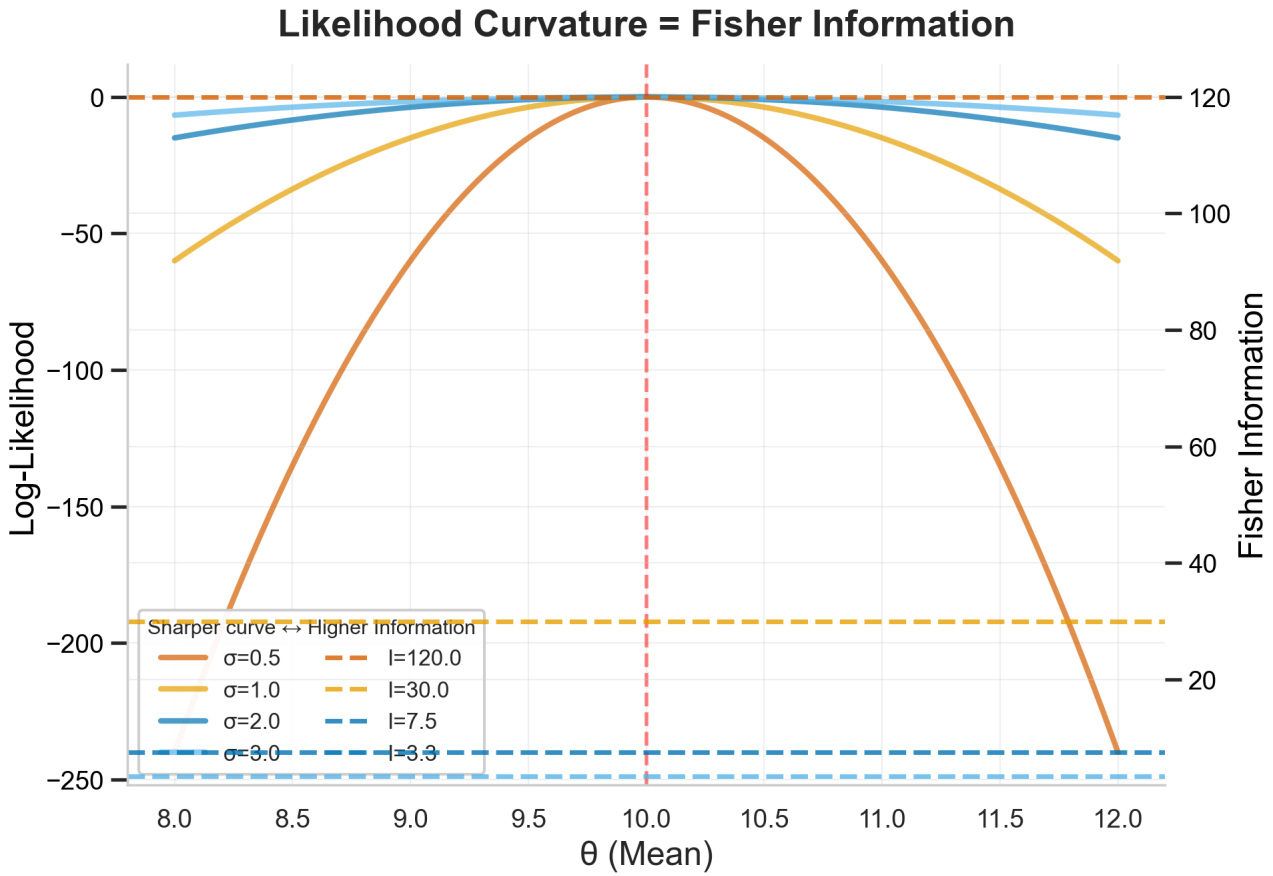
Takeaway: Smaller variance ⇒ sharper distribution ⇒ higher Fisher information.

Fisher Information — Visual (Poisson)



Takeaway: Fisher information decreases as the rate parameter increases for Poisson distributions.

Likelihood Curvature = Information



Takeaway: Sharper likelihood peaks correspond to higher Fisher information and lower achievable variance.

Cramér–Rao Lower Bound

Statement

For an unbiased estimator $\hat{\theta}$ of θ:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{\mathcal{I}_n(\theta)},$$

with equality if and only if $\hat{\theta}$ is the MLE (under regularity conditions).

Intuition

The CRLB quantifies the **minimum variance** achievable by any **unbiased** estimator, arising from the data's inherent variability and how sensitively the distribution depends on the parameter. No unbiased estimator can be more precise than this limit.

Practical Value

- Guides estimator selection during study design
- **Provides the foundation for MLE standard errors and CIs**
- Identifies when an estimator achieves optimal efficiency

Using the CRLB in Practice

Three Practical Uses

1. **Estimator Design (Theoretical):** Compare efficiency before data collection

$$\text{Efficiency}(\hat{\theta}) = \frac{\text{CRLB}}{\text{Var}(\hat{\theta})} = \frac{1/\mathcal{I}_n(\theta)}{\text{Var}(\hat{\theta})}$$

2. **MLE Standard Errors (Measurable):** Compute from data!

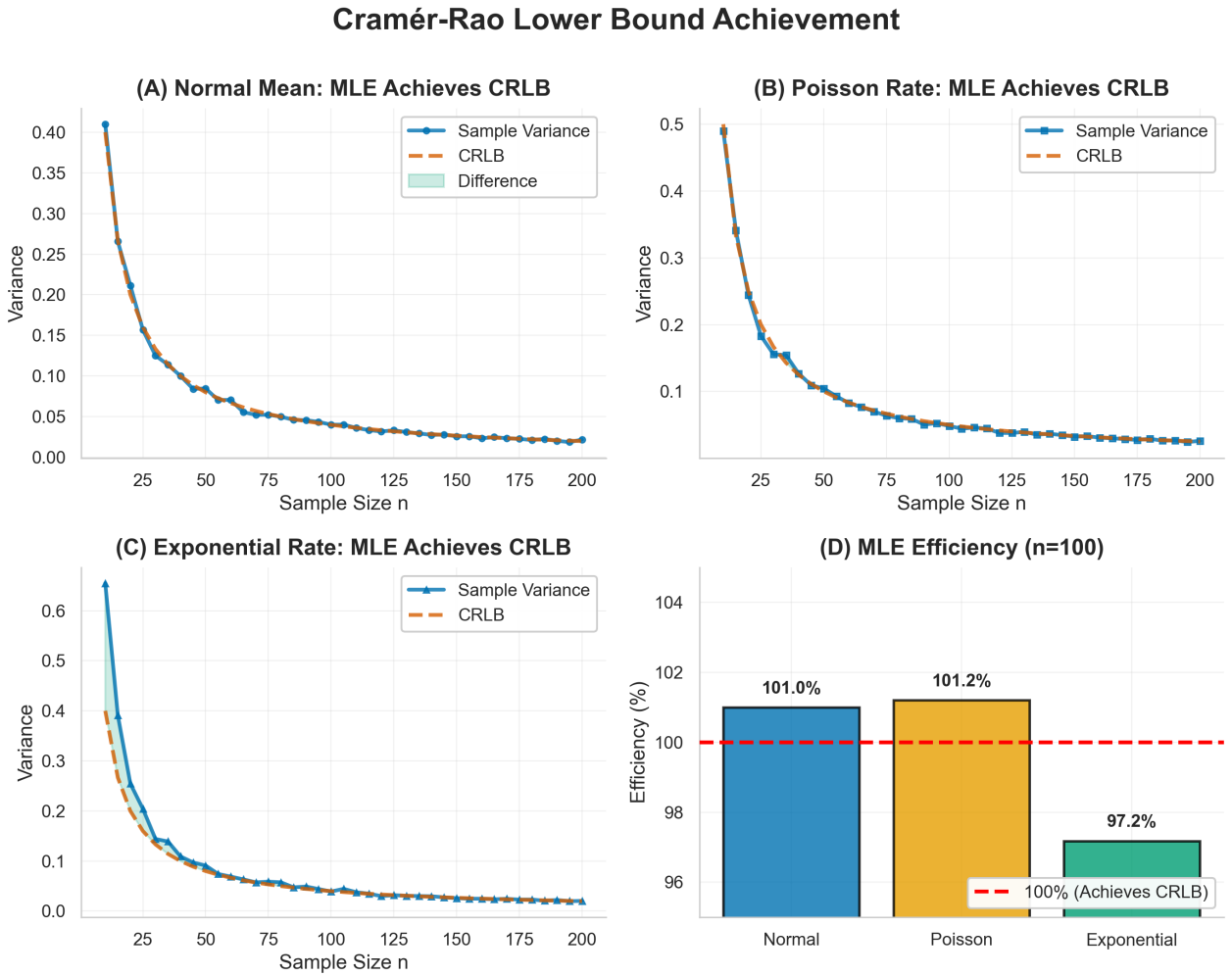
$$\widehat{\text{SE}}(\hat{\theta}_{\text{MLE}}) = \sqrt{\frac{1}{\hat{\mathcal{I}}_n(\hat{\theta})}} \quad \text{where} \quad \hat{\mathcal{I}}_n = -\frac{\partial^2 \ell}{\partial \theta^2}\bigg|_{\theta=\hat{\theta}}$$

3. **Method Comparison:** Estimate relative efficiency via bootstrap or asymptotic theory

Key Insight

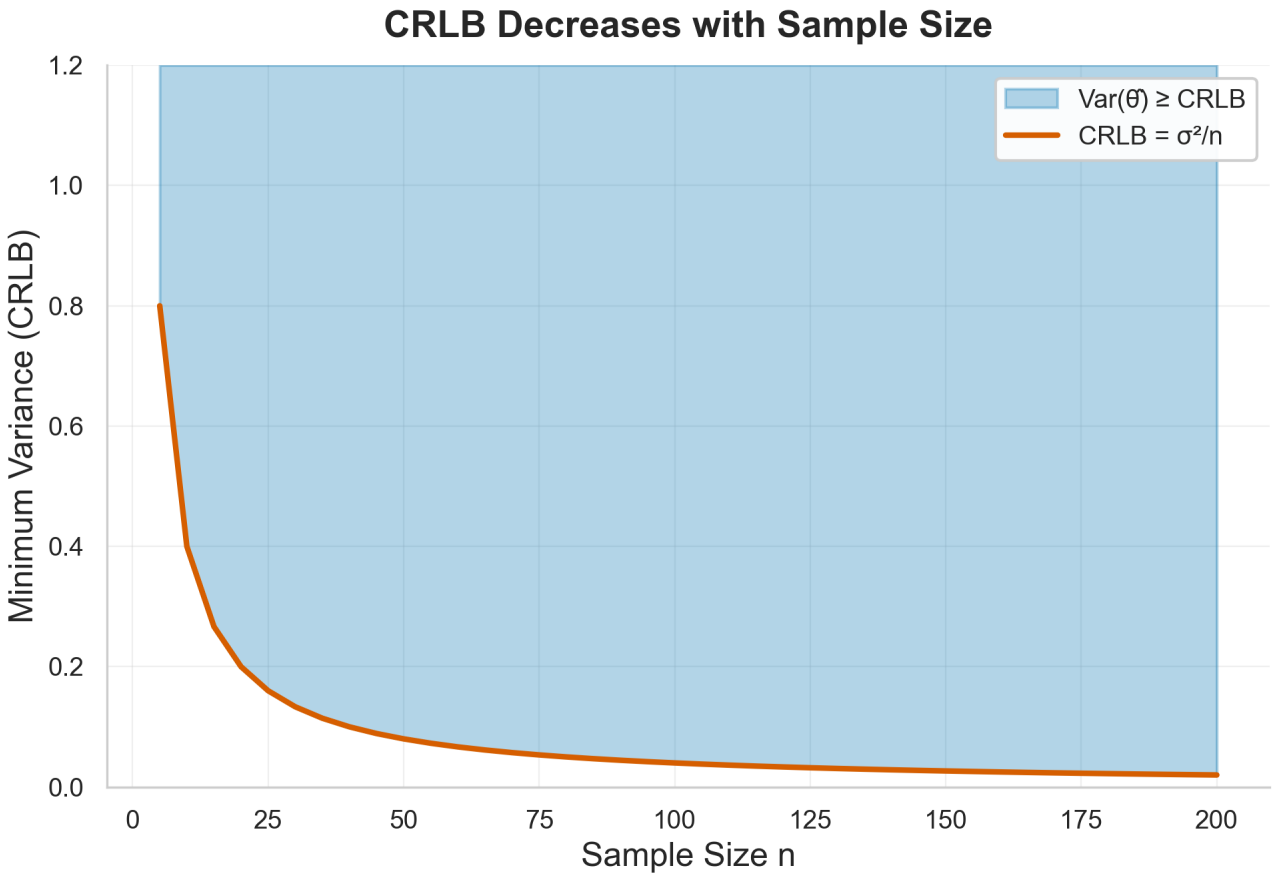
Every MLE confidence interval uses the CRLB! When you see "Std. Error" in software output for an MLE, that's the estimated CRLB.

Cramér–Rao Lower Bound — Visual



Takeaway: MLEs for Normal, Poisson, and Exponential achieve the CRLBmaximum efficiency.

CRLB vs Sample Size



Takeaway: The minimum achievable variance decreases as $1/n$, making larger samples more informative.

Example: Normal Mean (σ known)

Setup

$X_i \sim \mathcal{N}(\mu, \sigma^2)$ i.i.d. with σ known. Estimator: $\hat{\mu} = \bar{X}_n$.

Fisher Information

$$\mathcal{I}(\mu) = \frac{1}{\sigma^2}, \quad \mathcal{I}_n(\mu) = \frac{n}{\sigma^2}.$$

CRLB and Efficiency

$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n} = \frac{1}{\mathcal{I}_n(\mu)}.$$

The sample mean achieves the CRLB and is efficient.

Connection to Lesson 1

This follows from the CLT: $\sqrt{n}(\bar{X}_n - \mu) \to \mathcal{N}(0, \sigma^2)$.

Example: Poisson Rate

Setup

$X_i \sim \text{Poisson}(\lambda)$ i.i.d. Estimator: $\hat{\lambda} = \bar{X}_n$ (both MLE and MoM).

Fisher Information

$$\mathcal{I}(\lambda) = \frac{1}{\lambda}, \quad \mathcal{I}_n(\lambda) = \frac{n}{\lambda}.$$

CRLB and Efficiency

$$\text{Var}(\bar{X}_n) = \frac{\lambda}{n} = \frac{1}{\mathcal{I}_n(\lambda)}.$$

The sample mean achieves the CRLB and is efficient.

Connection to Lesson 2

This is the same estimator derived by both MLE and MoM methods, demonstrating efficiency of both approaches for this model.

Example: Exponential Rate

Setup

$X_i \sim \text{Exp}(\lambda)$ i.i.d. (rate parameter). MLE: $\hat{\lambda}_{\text{MLE}} = n / \sum_{i=1}^{n} X_i$.

Fisher Information

$$\mathcal{I}(\lambda) = \frac{1}{\lambda^2}, \quad \mathcal{I}_n(\lambda) = \frac{n}{\lambda^2}.$$

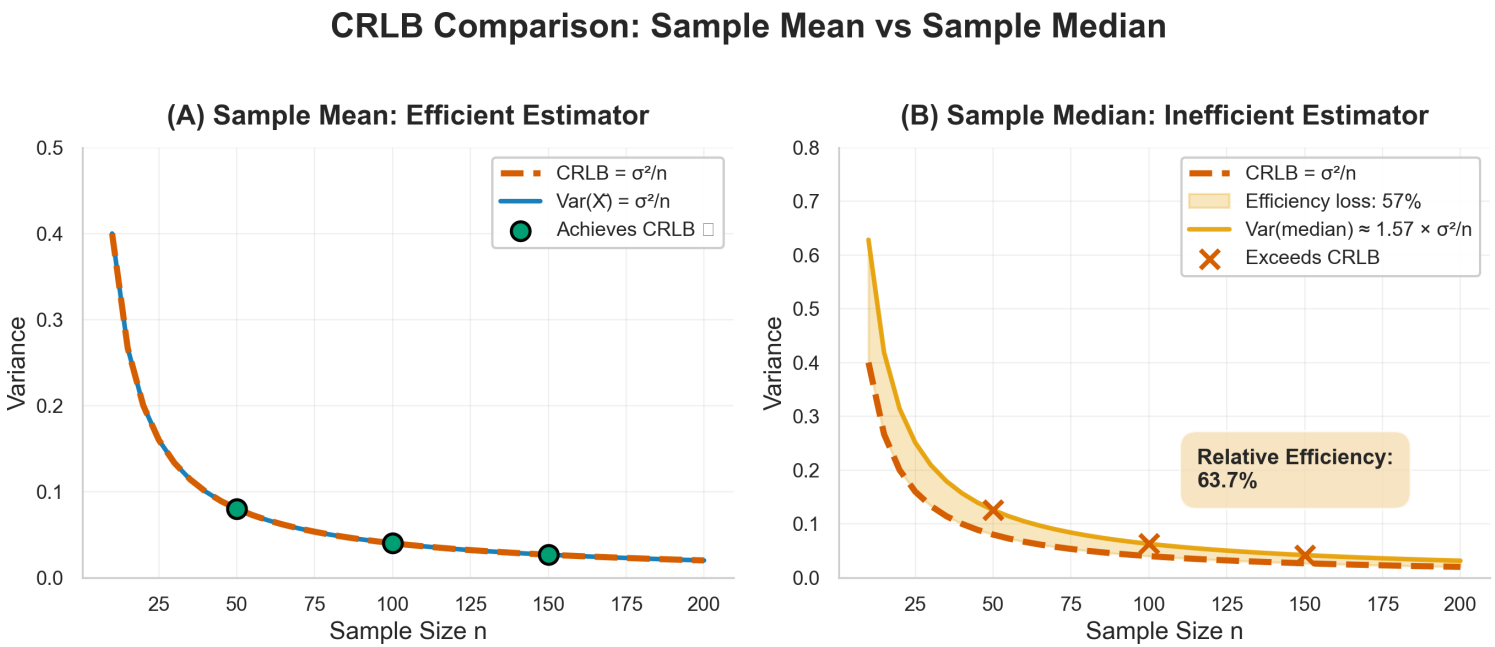
CRLB

$$\text{Var}(\hat{\lambda}) \geq \frac{\lambda^2}{n}.$$

Asymptotic Efficiency

The MLE achieves the CRLB asymptotically, but may have finite-sample bias.

Counter-Example: Sample Median for Normal Data



Key Insight

Both estimators are unbiased with the same CRLB $= \frac{\sigma^2}{n}$, but:

- **Sample mean:** Achieves CRLB—fully efficient ✓
- **Sample median:** Exceeds CRLB by 57%—wastes information

Takeaway: The CRLB helps identify which unbiased estimators use data most efficiently.

Asymptotic Normality of MLE

Statement

Under regularity conditions:

$$\sqrt{n}(\hat{\theta}_{\text{MLE}} - \theta_0) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathcal{I}(\theta_0)^{-1}\right).$$

Intuition

For large samples, MLEs behave like normal random variables with variance equal to the inverse Fisher information.

Practical Implication

This justifies the use of normal approximations for MLE confidence intervals and hypothesis tests in large samples.

Asymptotic Efficiency

Definition

An estimator is asymptotically efficient if it achieves the CRLB as $n \to \infty$, i.e.,

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \mathcal{I}(\theta)^{-1}\right).$$

MLE Property

Maximum likelihood estimators are asymptotically efficient under regularity conditions, making them the gold standard for large samples.

Connection to Lesson 2

This explains why MLEs are preferred when sample sizes are large and regularity conditions hold.

Pitfalls & Regularity Conditions

Regularity Conditions

- Support of distribution doesn't depend on θ
- Likelihood differentiable in θ
- Fisher information finite and positive
- Dominated convergence for expectation interchanges

When CRLB Fails

- Boundary parameters (e.g., variance near 0)
- Discrete parameters with small support
- Model misspecification

Exercises

1. Compute the Fisher information for Bernoulli(p) and derive the CRLB for unbiased estimators of p.
2. Show that the sample mean achieves the CRLB for Normal(μ, σ^2) with σ known.
3. Use the `fisher_info_poisson()` function from the appendix to compute information for $\lambda = 2, 5, 10$.
4. Explain why the MLE for Uniform$[0, \theta]$ achieves the CRLB while other estimators may not.

Summary

Key Takeaways

- Fisher information quantifies precision: $I(\theta) = \text{Var}(U(\theta))$
- CRLB sets minimum variance: $\text{Var}(\hat{\theta}) \geq 1/I_n(\theta)$
- MLEs achieve CRLB asymptotically (asymptotically efficient)
- Normal, Poisson, and Exponential examples illustrate efficiency
- Regularity conditions ensure CRLB applicability

Provides efficiency foundation for Lesson 2 estimators

- Wikipedia: Fisher information
  `https://en.wikipedia.org/wiki/Fisher_information`
- Casella & Berger, *Statistical Inference* (Chapter 7)

Outline

Confidence Intervals

Confidence Intervals

Learning Objectives

- Define $(1 - \alpha)$ confidence intervals and interpret correctly
- Derive classical CIs: Normal mean, t-intervals, variance, proportions
- Compare proportion CI methods using A/B testing data
- Understand pivots and asymptotic CIs (delta method)

Applies Lesson 1 (CLT) and Lesson 2 (delta method)

Confidence Interval Definition

Definition

A $(1-\alpha)$ confidence interval for θ is an interval $[L_n, U_n]$ such that

$$\mathbb{P}_\theta(L_n \leq \theta \leq U_n) = 1 - \alpha \quad \forall \theta.$$

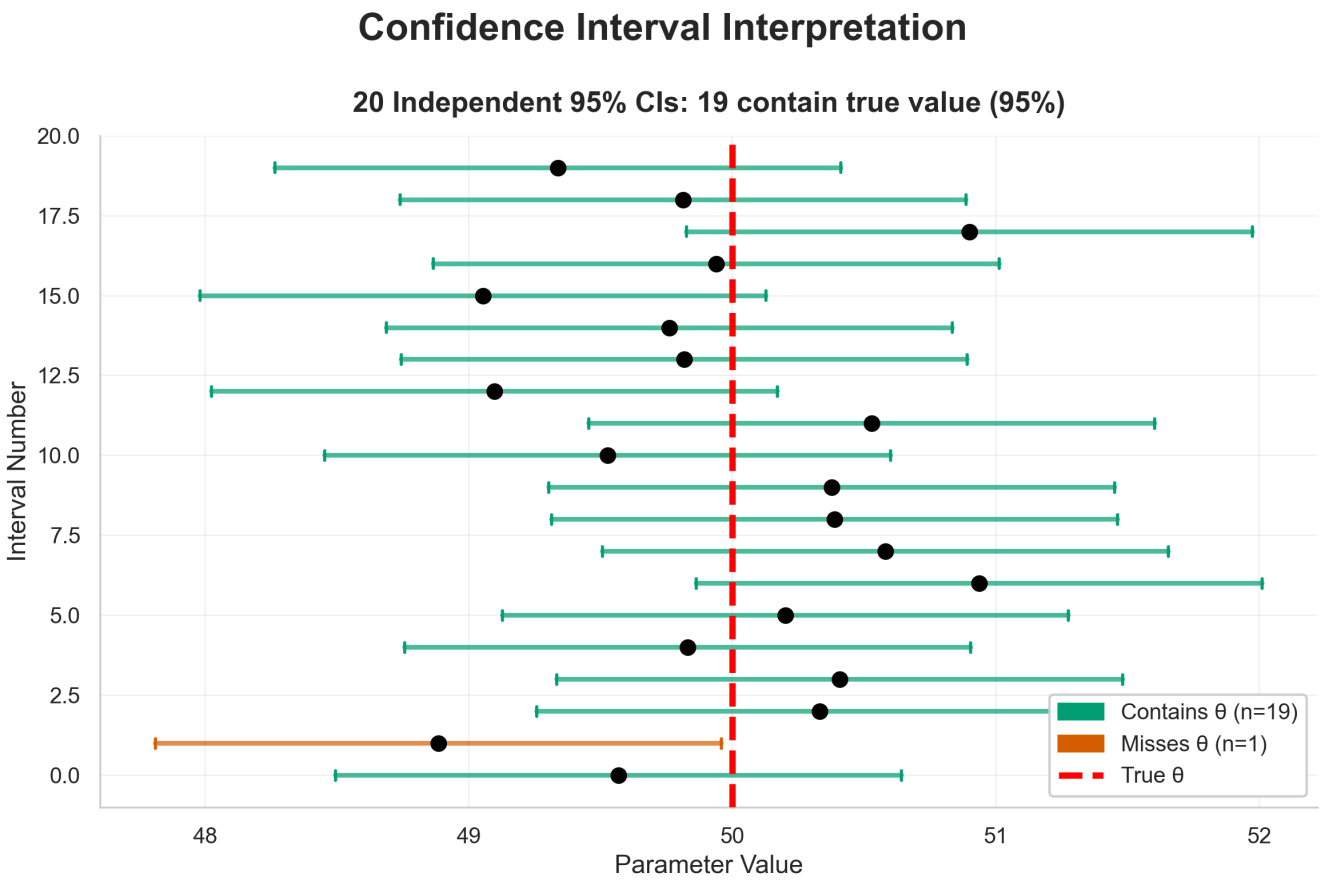
Interpretation

If we construct the interval many times, it will contain the true parameter in $(1-\alpha) \times 100\%$ of cases. This is a statement about the procedure, not any particular interval.

Common Misconception

A 95% CI does NOT mean there's a 95% probability that the true parameter is in *this particular* interval. The true parameter is fixed; the interval is random. We're 95% confident in the *procedure*, not in any single interval.

Confidence Interval — Visual



Pivotal Method

General Approach

A pivot is a function $P_n(\theta, X_1, \ldots, X_n)$ whose distribution doesn't depend on θ. Common pivots:

- $Z = \sqrt{n}(\bar{X}_n - \mu)/\sigma \sim \mathcal{N}(0,1)$
- $T = \sqrt{n}(\bar{X}_n - \mu)/S \sim t_{n-1}$
- $\chi^2 = (n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$

CI Construction

Find $q_{\alpha/2}, q_{1-\alpha/2}$ such that $\mathbb{P}(q_{\alpha/2} \leq P_n \leq q_{1-\alpha/2}) = 1 - \alpha$. Then solve for θ in the inequality.

Normal Mean CI (σ known)

Z-Interval

For $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ known:

$$\bar{X}_n \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Derivation

$$\mathbb{P}\left(\bar{X}_n - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X}_n + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha.$$

Connection to Lesson 1

This follows directly from the Central Limit Theorem and justifies normal-based inference for large samples.

Normal Mean CI (σ unknown)

t-Interval

For $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with σ unknown:

$$\bar{X}_n \pm t_{n-1,1-\alpha/2} \frac{S}{\sqrt{n}},$$

where $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$.

Pivot

The t-statistic $T = \sqrt{n}(\bar{X}_n - \mu)/S$ follows t_{n-1} distribution, accounting for uncertainty in the standard error estimate.

Small Sample Performance

t-intervals maintain nominal coverage even for small samples when normality holds, unlike normal approximations.

Normal Variance CI

Chi-Squared Interval

For $X_i \sim \mathcal{N}(\mu, \sigma^2)$ with μ unknown:

$$\left(\frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right).$$

Pivot

The chi-squared statistic $\chi^2 = (n-1)S^2/\sigma^2$ follows χ^2_{n-1} distribution, providing the basis for variance intervals.

Note

This interval is not symmetric around S^2 due to the skewness of the chi-squared distribution.

Binomial Proportion CIs

Wald Interval

$$\hat{p} \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Simple but poor coverage for small n or extreme p.

Wilson Score Interval

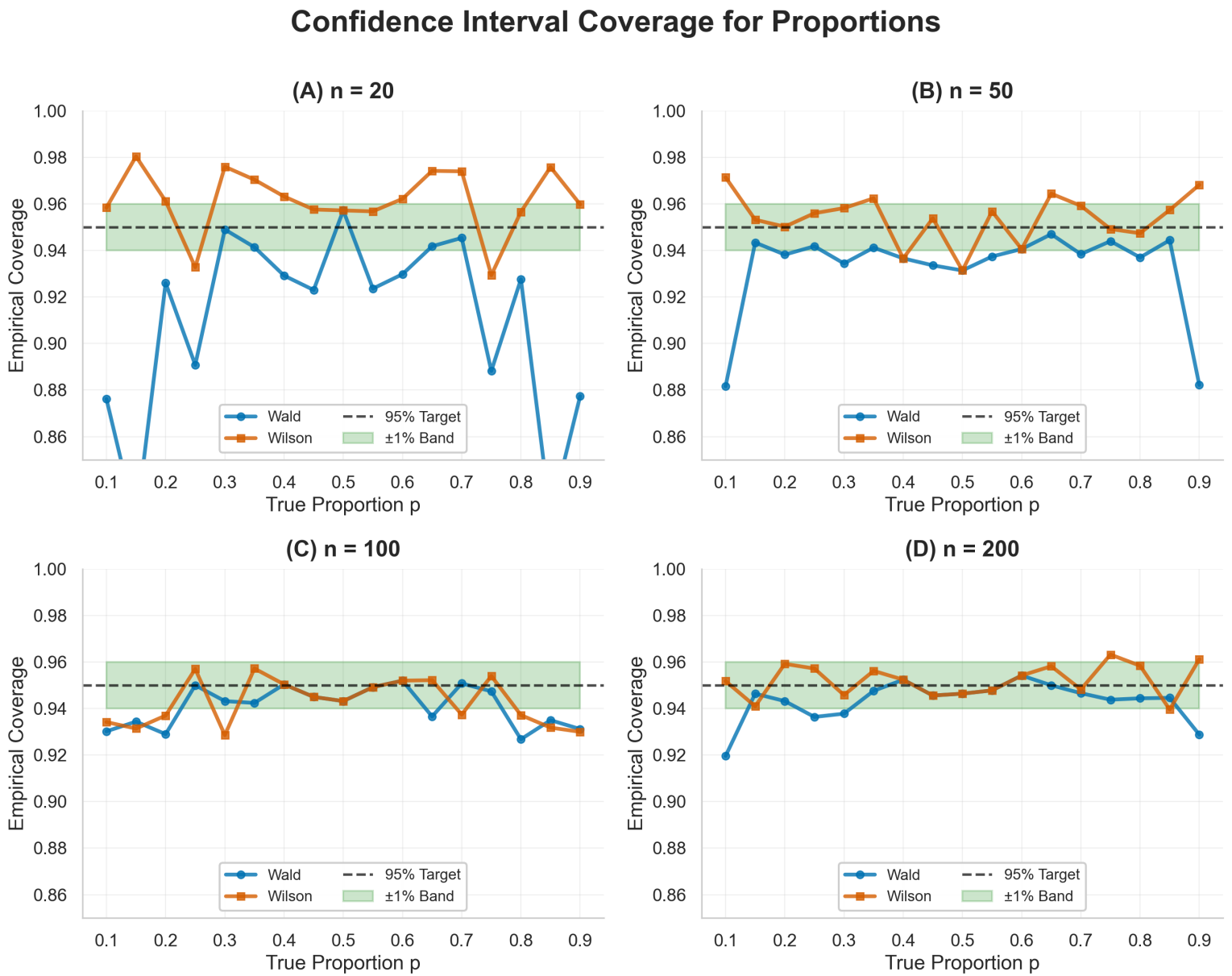
$$\frac{\hat{p} + z^2/(2n)}{1 + z^2/n} \pm \frac{z}{1 + z^2/n}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + z^2/(4n^2)}.$$

Better coverage but more complex formula.

Connection to A/B Testing

Use `shared/data/ab_test_clicks.csv` to compare methods with real data.

Proportion CI Comparison — Visual



Asymptotic CIs via Delta Method

Delta Method Statement

If $\sqrt{n}(\hat{\theta}_n - \theta) \to \mathcal{N}(0, \sigma^2)$, then for measurable function g,

$$\sqrt{n}(g(\hat{\theta}_n) - g(\theta)) \to \mathcal{N}(0, [g'(\theta)]^2\sigma^2).$$

CI Construction

Approximate CI for $g(\theta)$:

$$g(\hat{\theta}_n) \pm z_{1-\alpha/2}\sqrt{[g'(\hat{\theta}_n)]^2\widehat{\text{Var}}(\hat{\theta}_n)/n}.$$

Connection to Lesson 2

The delta method extends asymptotic normality of MLEs to functions of parameters, enabling inference for ratios, logs, etc.

Delta Method Example: Log-Odds

Problem

Estimate CI for log-odds $\theta = \log\left(\frac{p}{1-p}\right)$ from $\hat{p} = X/n$ where $X \sim \text{Binomial}(n, p)$.

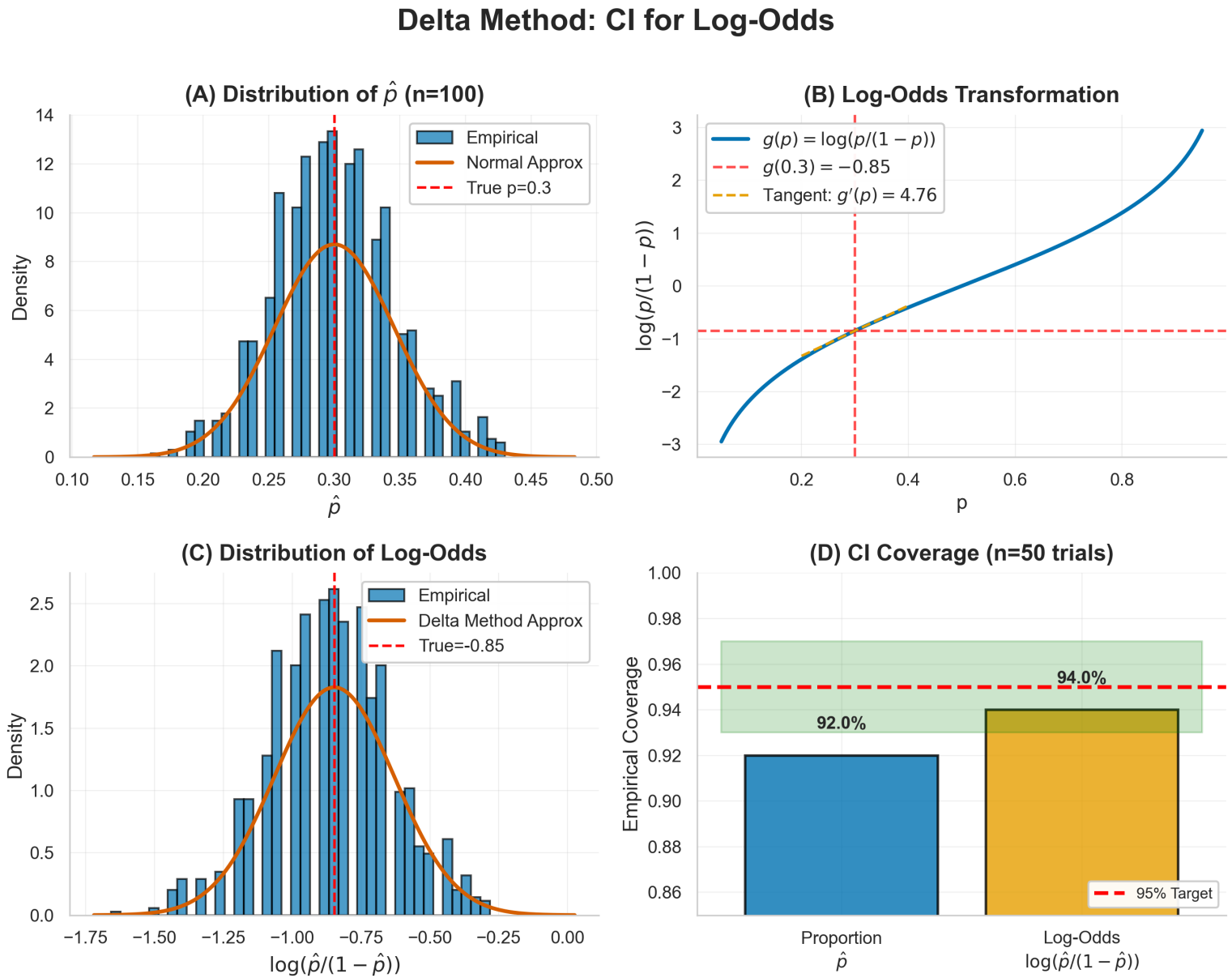
Solution Steps

1. For proportion $\hat{p}$: $\sqrt{n}(\hat{p} - p) \to \mathcal{N}(0, p(1-p))$
2. Transformation: $g(p) = \log\left(\frac{p}{1-p}\right)$
3. Derivative: $g'(p) = \frac{1}{p(1-p)}$
4. Delta method: $\sqrt{n}(g(\hat{p}) - g(p)) \to \mathcal{N}\left(0, \frac{1}{p(1-p)}\right)$

95% CI for Log-Odds

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) \pm 1.96\sqrt{\frac{1}{n\hat{p}(1-\hat{p})}}.$$

Delta Method — Visual



Practical Advice

When to Use Each Method

- Z-intervals: Large samples, known variance
- t-intervals: Small samples, unknown variance, normality
- Wilson intervals: Proportions, especially small n or extreme p
- Delta method: Functions of parameters, large samples

General Guidelines

- Check assumptions (normality, sample size)
- Compare interval width vs coverage
- Use simulation to verify performance
- Consider bootstrap for complex scenarios

Exercises

1. Derive the t-interval for Normal mean with unknown variance using the pivotal method.
2. Show that the Wilson interval for p=0.5 and large n reduces to the Wald interval.
3. Use A/B testing data from `shared/data/ab_test_clicks.csv` to compute and compare proportion CIs.
4. Apply the delta method to construct a CI for the coefficient of variation σ/μ.

Summary

Key Takeaways

- CIs quantify uncertainty in parameter estimates
- Pivotal method provides general CI construction framework
- t-intervals improve on normal approximations for small samples
- Wilson intervals provide better proportion coverage than Wald
- Delta method enables inference for parameter functions

Foundation for hypothesis testing and decision making

- Wikipedia: Confidence interval
  `https://en.wikipedia.org/wiki/Confidence_interval`
- Casella & Berger, *Statistical Inference* (Chapter 9)

Outline

Bootstrap

Bootstrap

Learning Objectives

- Explain the bootstrap idea: resampling to approximate sampling distributions
- Implement percentile and basic bootstrap confidence intervals
- Apply bootstrap to non-smooth statistics (median, quantiles)
- Compare bootstrap CI coverage to parametric methods

Complements confidence intervals, builds on Lesson 1 & 2

Motivation

When Analytic Methods Fail

- Complex statistics (medians, quantiles, correlations)
- Non-standard distributions or small samples
- Model misspecification or unknown sampling distributions
- Functions of multiple parameters

Bootstrap Idea

Use the data itself as an estimate of the population distribution. Resample with replacement to approximate the sampling distribution of any statistic of interest.

Connection to Lesson 1

Bootstrap approximates the sampling distribution without knowing the true population parameters or distribution form.

Nonparametric Bootstrap Algorithm

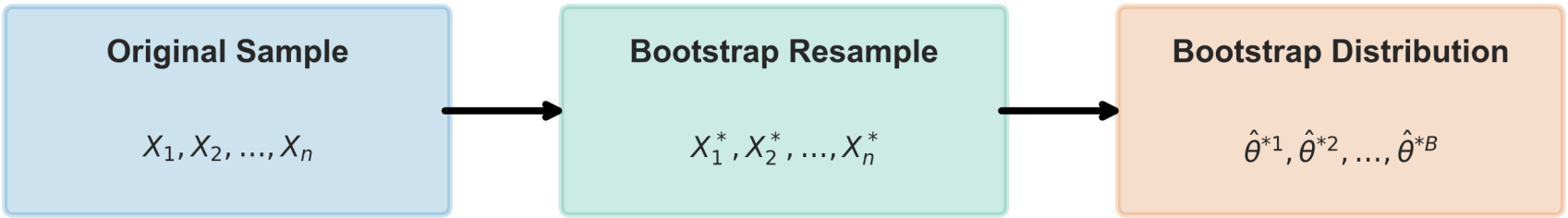
Steps

1. Compute statistic of interest: $\hat{\theta} = g(X_1, \ldots, X_n)$
2. For $b = 1$ to B:
   - Draw bootstrap sample $X_1^*, \ldots, X_n^* \sim$ empirical distribution
   - Compute $\hat{\theta}^{*b} = g(X_1^{*b}, \ldots, X_n^{*b})$
3. Use $\{\hat{\theta}^{*1}, \ldots, \hat{\theta}^{*B}\}$ for inference

Key Insight

The empirical distribution $\hat{F}_n$ serves as a nonparametric estimate of the true distribution F, enabling resampling without parametric assumptions.

Bootstrap Algorithm – Visual



Resample with replacement

from empirical distribution

Bootstrap Confidence Intervals

Percentile Interval

$$\left[q^*_{\alpha/2}, q^*_{1-\alpha/2}\right],$$

where q^*_p is the p-quantile of bootstrap replicates.

Basic Interval (Reflection Method)

$$\hat{\theta} - (q^*_{1-\alpha/2} - \hat{\theta}) \quad \text{to} \quad \hat{\theta} + (\hat{\theta} - q^*_{\alpha/2})$$

Reflects the bootstrap quantiles around $\hat{\theta}$ to center the interval.
Equivalently: $\left[2\hat{\theta} - q^*_{1-\alpha/2}, 2\hat{\theta} - q^*_{\alpha/2}\right]$.

BCa Interval (Advanced)

Bias-corrected and accelerated interval that adjusts for bias and skewness in the bootstrap distribution.

Example: Median CI (Exponential Data)

Setup

$X_i \sim \text{Exp}(\lambda)$ i.i.d. (skewed distribution). Statistic: sample median $\hat{\theta} = \text{median}(X_1, \ldots, X_n)$.

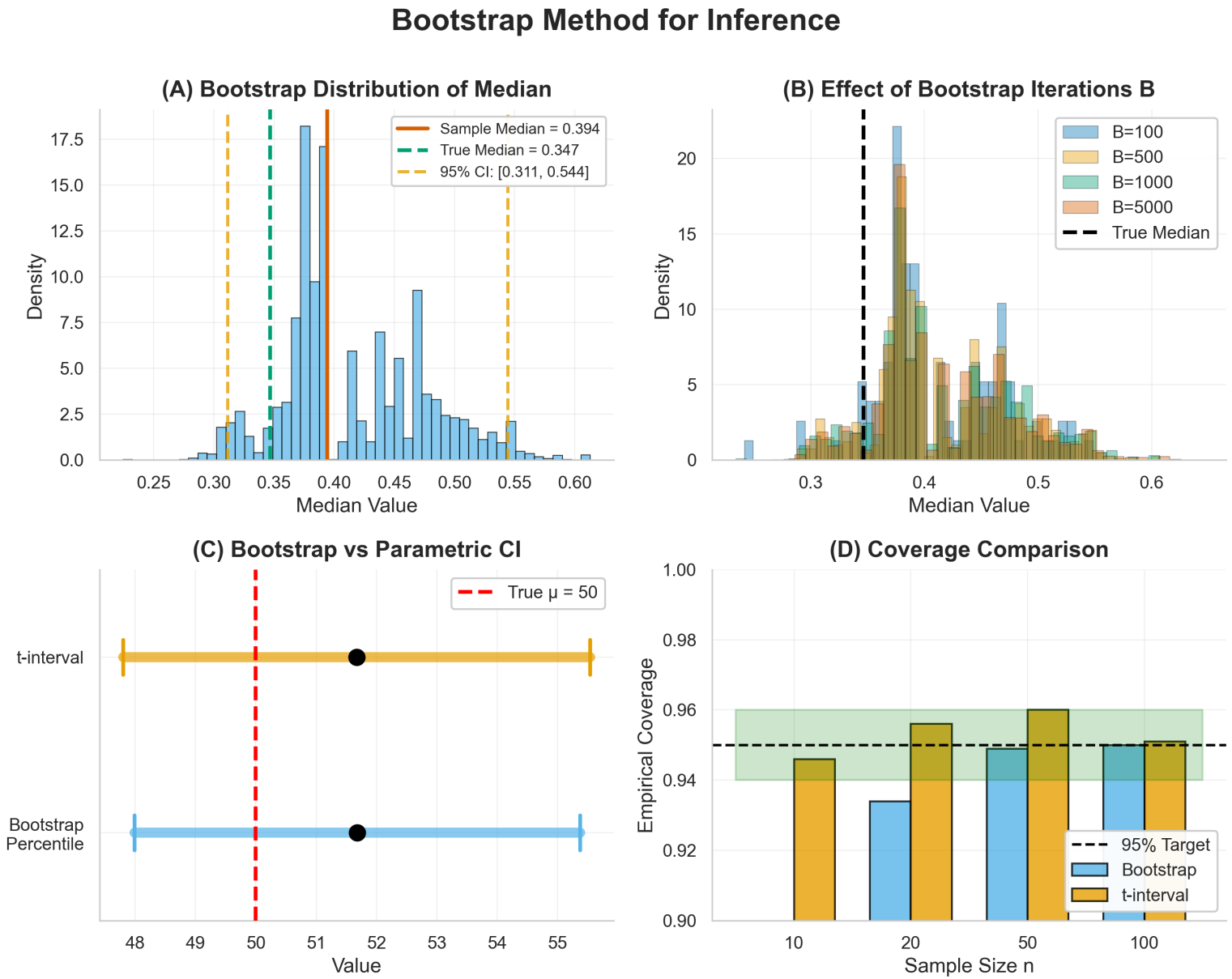
Challenge

The sampling distribution of the median is complex and depends on the underlying distribution, making analytic CIs difficult.

Bootstrap Solution

Use nonparametric bootstrap to approximate the sampling distribution of the median and construct percentile or basic intervals.

Median Bootstrap — Visual



Example: Difference in Means

A/B Testing Scenario

Compare means between two groups with potentially different variances. Use heights data from `shared/data/heights_weights_sample.csv`.

Challenge

Welch's t-test assumes normality and provides analytic intervals, but bootstrap offers a robust alternative without strong assumptions.

Bootstrap Approach

Bootstrap both groups separately and compute bootstrap distribution of the difference in means for inference.

Practical Considerations

Number of Bootstrap Resamples

- $B = 1000$ often sufficient for basic intervals
- $B = 5000$ or more for precise quantile estimation
- Computational cost scales with $B \times n$
- Parallelization can speed up computation

Random Seeds

Always set random seeds for reproducibility:

- Different seeds can give slightly different results
- Report seeds in publications and assignments
- Use `rng = np.random.default_rng(2025)`

BCa Method (Conceptual)

Bias Correction

Adjusts for bias in the bootstrap distribution when the statistic is not centered at the true parameter.

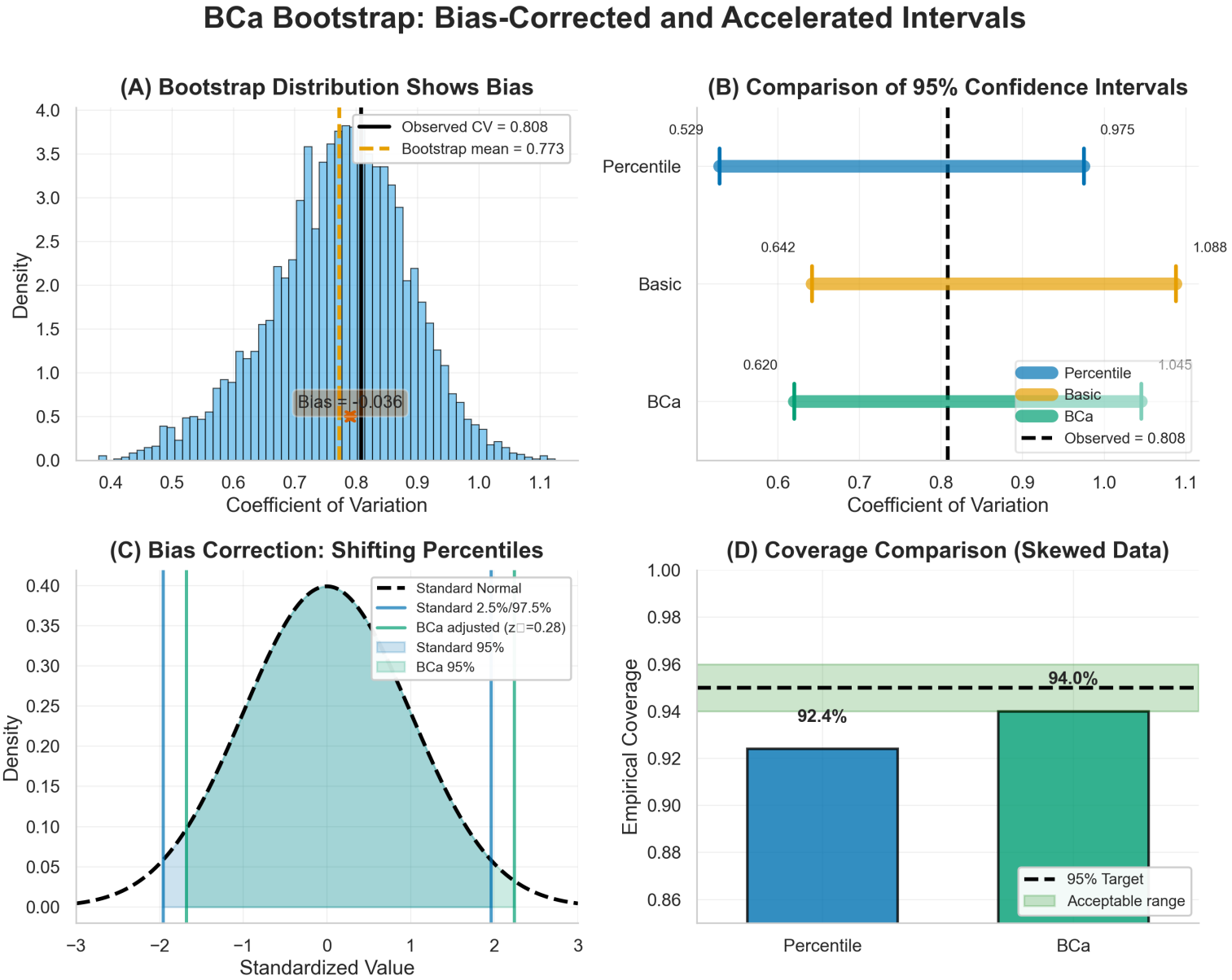
Acceleration

Adjusts for skewness and heteroscedasticity in the bootstrap distribution using jackknife estimates.

When Helpful

BCa intervals often provide better coverage than percentile intervals, especially for skewed distributions or small samples.

BCa Method – Visual Example



Pitfalls and Limitations

When Bootstrap Fails

- Dependent data (requires block bootstrap or other methods)
- Very small samples ($n < 10$) may not work well
- Heavy-tailed distributions may need many resamples
- Boundary parameters (variances, correlations) need care

Best Practices

- Always compare to parametric methods when available
- Check bootstrap distribution shape for anomalies
- Use multiple random seeds to assess stability
- Consider parametric bootstrap when model is trusted

Exercises

1. Use bootstrap to construct a 95% confidence interval for the median of exponential data.
2. Compare bootstrap CI for mean difference vs Welch's t-interval using heights data from `shared/data/heights_weights_sample.csv`.
3. Implement studentized bootstrap for the sample mean and compare to percentile bootstrap.
4. Investigate how bootstrap performance degrades with very small sample sizes ($n = 5, 10$).

Summary

Key Takeaways

- Bootstrap approximates sampling distributions via resampling
- Percentile and basic intervals are most common
- Bootstrap excels for complex statistics (medians, quantiles)
- Provides robust alternative to parametric methods
- Requires careful consideration of sample size and dependence

Robust complement to parametric confidence intervals

- Wikipedia: Bootstrap (statistics)
  `https://en.wikipedia.org/wiki/Bootstrapping_(statistics)`
- Efron & Tibshirani, *An Introduction to the Bootstrap*