Lecture 8: Confidence Intervals & the Bootstrap

Two Paths to Quantifying Uncertainty

Previously, on Lecture 7...

Sampling distribution: $\hat{\theta}$ is random — different samples give different estimates.

CLT: $\bar{X} \dot\sim N(\mu, \sigma^2/n)$. The MLE is also asymptotically Normal: $\hat{\theta} \dot\sim N(\theta, 1/(nI(\theta)))$.

Standard error: $\mathrm{SE} = \mathrm{SD}(\hat{\theta})$. For $\bar{X}$: $\mathrm{SE} = \sigma/\sqrt{n}$. For any MLE: $\mathrm{SE} \approx 1/\sqrt{nI(\hat{\theta})}$.

$\sqrt{n}$ **law:** Halving the SE requires quadrupling n. Precision is expensive.

SD $\neq$ SE: SD measures data spread (fixed). SE measures estimator precision (shrinks with n).

Today: Two ways to **use** the sampling distribution.

"52% support candidate A" means nothing without $\pm$ something.

Part I: Confidence Intervals

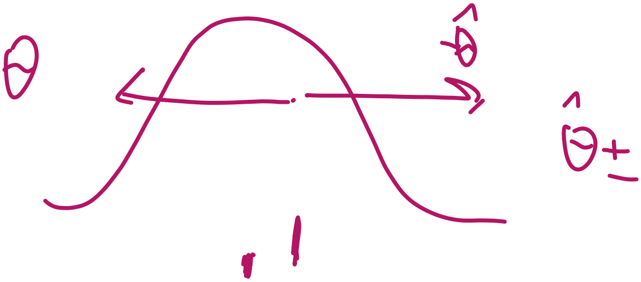
The analytical path: formulas from theory

What Is a Confidence Interval?

A **95% confidence interval** is a random interval $[L, U]$ such that

$$P(L \leq \theta \leq U) = 0.95$$

Before you collect data, there is a 95% chance the interval will contain θ.

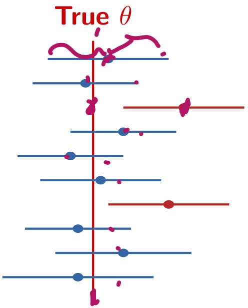


What Is a Confidence Interval?

A **95% confidence interval** is a random interval $[L, U]$ such that

$$P(L \le \theta \le U) = 0.95$$

Before you collect data, there is a 95% chance the interval will contain θ.



10 simulated 95% CIs
2 miss θ (expected $\approx$ 1 in 20)

From Sampling Distribution to Interval

The problem: A point estimate $\hat{\theta}$ is never exactly right. We need to report a **range of plausible values**.

From Sampling Distribution to Interval

The problem: A point estimate $\hat{\theta}$ is never exactly right. We need to report a **range of plausible values**.

Key insight: From Lecture 7, we know $\hat{\theta}$ is approximately Normal: $\frac{\hat{\theta}-\theta}{\text{SE}(\hat{\theta})} \approx N(0,1)$.

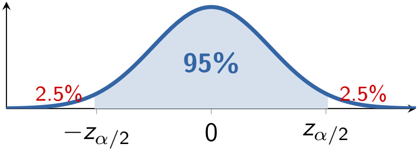
So with 95% probability, this standardized quantity falls between ± 1.96. Rearranging for θ...

From Sampling Distribution to Interval

The problem: A point estimate $\hat{\theta}$ is never exactly right. We need to report a **range of plausible values**.

Key insight: From Lecture 7, we know $\hat{\theta}$ is approximately Normal: $\frac{\hat{\theta}-\theta}{\mathrm{SE}(\hat{\theta})} \approx N(0,1)$.

So with 95% probability, this standardized quantity falls between ± 1.96. Rearranging for θ...

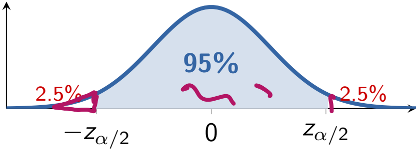


From Sampling Distribution to Interval

The problem: A point estimate $\hat{\theta}$ is never exactly right. We need to report a **range of plausible values**.

Key insight: From Lecture 7, we know $\hat{\theta}$ is approximately Normal: $\frac{\hat{\theta}-\theta}{\text{SE}(\hat{\theta})} \approx N(0,1)$.

So with 95% probability, this standardized quantity falls between ± 1.96. Rearranging for θ...



The Wald CI (named after Abraham Wald, 1943):

$$\hat{\theta} \quad \pm \quad z_{\alpha/2} \cdot \text{SE}(\hat{\theta})$$

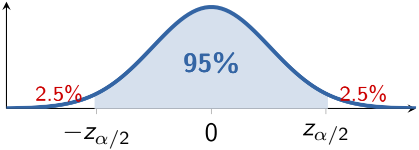
For 95%: $\hat{\theta} \pm 1.96 \cdot \text{SE}$ For 99%: $\hat{\theta} \pm 2.576 \cdot \text{SE}$ For 90%: $\hat{\theta} \pm 1.645 \cdot \text{SE}$

From Sampling Distribution to Interval

The problem: A point estimate $\hat{\theta}$ is never exactly right. We need to report a **range of plausible values**.

Key insight: From Lecture 7, we know $\hat{\theta}$ is approximately Normal: $\frac{\hat{\theta}-\theta}{\text{SE}(\hat{\theta})} \approx N(0,1)$.

So with 95% probability, this standardized quantity falls between ± 1.96. Rearranging for θ...



The Wald CI (named after Abraham Wald, 1943):

$$\hat{\theta} \quad \pm \quad z_{\alpha/2} \cdot \text{SE}(\hat{\theta})$$

For 95%: $\hat{\theta} \pm 1.96 \cdot \text{SE}$ For 99%: $\hat{\theta} \pm 2.576 \cdot \text{SE}$ For 90%: $\hat{\theta} \pm 1.645 \cdot \text{SE}$

Recipe:

1. Compute the point estimate $\hat{\theta}$
2. Compute (or estimate) the standard error $\text{SE}(\hat{\theta})$

Example: CI for the Mean

Setup: $n = 36$ lightbulbs tested. $\bar{X} = 1{,}200$ hours, $S = 120$ hours.

Example: CI for the Mean

Setup: $n = 36$ lightbulbs tested. $\bar{X} = 1{,}200$ hours, $S = 120$ hours.

Step 1: Point estimate: $\hat{\mu} = \bar{X} = 1{,}200$.

Example: CI for the Mean

Setup: $n = 36$ lightbulbs tested. $\bar{X} = 1{,}200$ hours, $S = 120$ hours.

Step 1: Point estimate: $\hat{\mu} = \bar{X} = 1{,}200$.
Step 2: Standard error: $\widehat{\text{SE}} = S/\sqrt{n} = 120/\sqrt{36} = 20$.

$$\frac{\sigma}{\sqrt{n}}$$

Example: CI for the Mean

Setup: $n = 36$ lightbulbs tested. $\bar{X} = 1{,}200$ hours, $S = 120$ hours.

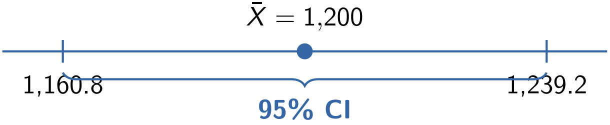
Step 1: Point estimate: $\hat{\mu} = \bar{X} = 1{,}200$.
Step 2: Standard error: $\widehat{\mathrm{SE}} = S/\sqrt{n} = 120/\sqrt{36} = 20$.
Step 3: For 95%: $z_{0.025} = 1.96$.

Example: CI for the Mean

Setup: $n = 36$ lightbulbs tested. $\bar{X} = 1{,}200$ hours, $S = 120$ hours.

Step 1: Point estimate: $\hat{\mu} = \bar{X} = 1{,}200$.
Step 2: Standard error: $\widehat{\text{SE}} = S/\sqrt{n} = 120/\sqrt{36} = 20$.
Step 3: For 95%: $z_{0.025} = 1.96$.
Step 4:

$$\bar{X} \pm 1.96 \cdot \text{SE} = 1{,}200 \pm 1.96 \times 20 = 1{,}200 \pm 39.2$$



"We are 95% confident that the true mean lifetime is between 1,161 and 1,239 hours."

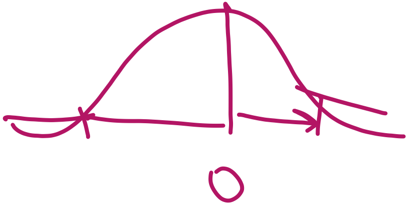
Standard format in papers: "Mean lifetime was 1,200 hours (95% CI: 1,161–1,239)."

What 95% Confidence **Really** Means

Correct:

If we repeated the experiment many times, 95% of the resulting intervals would contain θ.

The *procedure* works 95% of the time.



Simulation check: build 10,000 CIs from $N(\mu, \sigma^2)$ — about 9,500 contain μ. You'll verify this in the practical.

Analogy: A 95%-accurate archer hits the target 95% of the time. After the arrow lands, it either hit or missed — the probability is gone.

A CI is the arrow. Once computed, it either contains θ or it doesn't.

What 95% Confidence **Really** Means

Correct:

If we repeated the experiment many times, 95% of the resulting intervals would contain θ.

The *procedure* works 95% of the time.

Common mistake:

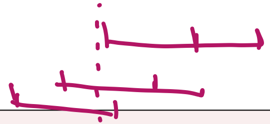
"There is a 95% probability that θ lies in $[L, U]$."

θ is fixed! It's either in there or not.

The randomness is in the *interval*, not in θ.

Simulation check: build 10,000 CIs from $N(\mu, \sigma^2)$ — about 9,500 contain μ. You'll verify this in the practical.

What 95% Confidence **Really** Means



Correct:

If we repeated the experiment many times, 95% of the resulting intervals would contain θ.

The *procedure* works 95% of the time.

Common mistake:

"There is a 95% probability that θ lies in $[L, U]$."

θ is fixed! It's either in there or not. The randomness is in the *interval*, not in θ.

Simulation check: build 10,000 CIs from $N(\mu, \sigma^2)$ — about 9,500 contain μ. You'll verify this in the practical.

Analogy: A 95%-accurate archer hits the target 95% of the time. After the arrow lands, it either hit or missed — the probability is gone.

A CI is the arrow. Once computed, it either contains θ or it doesn't.

Quick Check: Can You Say It Right?

A study reports: "Mean recovery time was 12.3 days (95% CI: 10.1–14.5)."

Which interpretations are **correct**?

✗ "There is a 95% probability that μ is between 10.1 and 14.5."
μ is fixed. After computing, it's either in there or not. The probability is about the *method*.

Quick Check: Can You Say It Right?

A study reports: "Mean recovery time was 12.3 days (95% CI: 10.1–14.5)."

Which interpretations are **correct**?

✗ "There is a 95% probability that μ is between 10.1 and 14.5."
μ is fixed. After computing, it's either in there or not. The probability is about the *method*.

✓ "We are 95% confident that μ lies between 10.1 and 14.5 days."
"95% confident" = if we repeated, 95% of intervals would capture μ.

Quick Check: Can You Say It Right?

A study reports: "Mean recovery time was 12.3 days (95% CI: 10.1–14.5)."

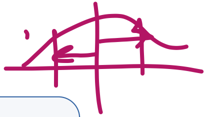
Which interpretations are **correct**?

✗ "There is a 95% probability that μ is between 10.1 and 14.5."
μ is fixed. After computing, it's either in there or not. The probability is about the *method*.

✓ "We are 95% confident that μ lies between 10.1 and 14.5 days."
"95% confident" = if we repeated, 95% of intervals would capture μ.

✗ "95% of patients recover in 10.1 to 14.5 days."
Confuses a CI for the *mean* with a *prediction interval* for individuals.

What Determines the Width?



$$\text{Width} = 2 \cdot z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Confidence level $\uparrow$

$z_{\alpha/2}$ increases

$\Rightarrow$ **wider** CI

90%: 1.645
95%: 1.960
99%: 2.576

What Determines the Width?

$$\text{Width} = 2 \cdot z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Confidence level $\uparrow$

$z_{\alpha/2}$ increases

$\Rightarrow$ **wider** CI

90%: 1.645
95%: 1.960
99%: 2.576

Variability σ $\uparrow$

More noise

$\Rightarrow$ **wider** CI

Noisier data means less precision

What Determines the Width?

$$\text{Width} = \quad 2 \cdot z_{\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

Confidence level $\uparrow$

$z_{\alpha/2}$ increases

$\Rightarrow$ **wider** CI

90%: 1.645
95%: 1.960
99%: 2.576

Variability $\sigma \uparrow$

More noise

$\Rightarrow$ **wider** CI

Noisier data means less precision

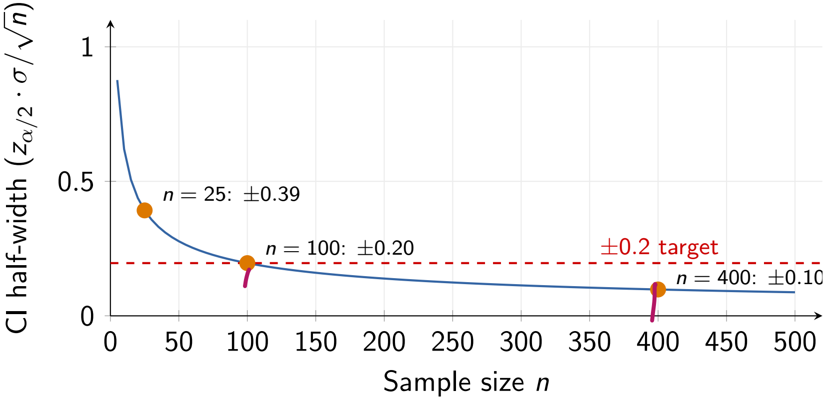
Sample size $n \uparrow$

$\sqrt{n}$ in denominator

$\Rightarrow$ **narrower** CI

But diminishing returns ($\sqrt{n}$ law!)

The $\sqrt{n}$ Law in Action



Halving the width requires **quadrupling** n. From $n = 100$ to ± 0.10 needs $n = 400$.

(Shown for $\sigma = 1$, 95% confidence.)

CI for a Proportion

Setup: $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. MLE: $\hat{p} = k/n$. SE $= \sqrt{\hat{p}(1-\hat{p})/n}$.

CI for a Proportion

Setup: $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. MLE: $\hat{p} = k/n$. SE $= \sqrt{\hat{p}(1-\hat{p})/n}$.

Wald 95% CI for a proportion:

$$\hat{p} \qquad \pm \qquad 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

CI for a Proportion

Setup: $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. MLE: $\hat{p} = k/n$. SE $= \sqrt{\hat{p}(1-\hat{p})/n}$.

Wald 95% CI for a proportion:

$$\hat{p} \qquad \pm \qquad 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Example: $n = 1{,}000$, $\hat{p} = 0.52$.

SE $= \sqrt{0.52 \times 0.48/1000} = 0.0158$

95% CI: $0.52 \pm 1.96 \times 0.0158 = [0.489,\ 0.551]$.

CI for a Proportion

Setup: $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. MLE: $\hat{p} = k/n$. SE $= \sqrt{\hat{p}(1-\hat{p})/n}$.

Wald 95% CI for a proportion:

$$\hat{p} \qquad \pm \qquad 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\hat{p} = \frac{k}{n}$

Example: $n = 1{,}000$, $\hat{p} = 0.52$.

SE $= \sqrt{0.52 \times 0.48/1000} = 0.0158$

95% CI: $0.52 \pm 1.96 \times 0.0158 = [0.489,\ 0.551]$.

Problem: $n = 20$, $k = 0$. Then $\hat{p} = 0$ and SE $= 0$.

Wald gives $[0,\ 0]$. **Useless!**

We *know* p could be positive — zero observations doesn't mean zero probability.

The Wilson Fix

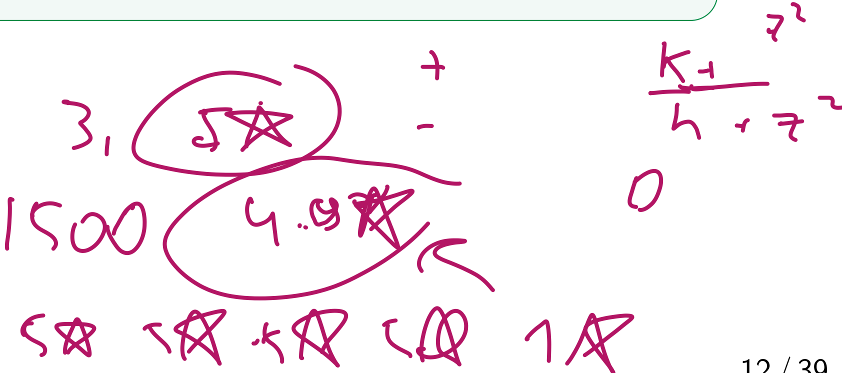
The Wald CI collapses at the boundary. The **Wilson interval** fixes this by adding "pseudo-observations" (like MAP!):

The Wilson Fix

The Wald CI collapses at the boundary. The **Wilson interval** fixes this by adding "pseudo-observations" (like MAP!):

$$\tilde{p} = \frac{k+z^2/2}{n+z^2} \qquad \text{Wilson CI:} \quad \tilde{p} \pm \frac{z}{n+z^2}\sqrt{k(n-k)/n + z^2/4}$$

For 95%: acts like adding 2 successes and 2 failures $\Rightarrow \tilde{p} \approx (k+2)/(n+4)$.



The Wilson Fix

The Wald CI collapses at the boundary. The **Wilson interval** fixes this by adding "pseudo-observations" (like MAP!):

$$\tilde{p} = \frac{k+z^2/2}{n+z^2} \qquad \text{Wilson CI:} \quad \tilde{p} \pm \frac{z}{n+z^2}\sqrt{k(n-k)/n + z^2/4}$$

For 95%: acts like adding 2 successes and 2 failures $\Rightarrow \tilde{p} \approx (k+2)/(n+4)$.

Rule of thumb: Use Wald when $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.
Otherwise use Wilson. Many stats packages default to Wilson.

The Wilson Fix

The Wald CI collapses at the boundary. The **Wilson interval** fixes this by adding "pseudo-observations" (like MAP!):

$$\tilde{p} = \frac{k+z^2/2}{n+z^2} \qquad \text{Wilson CI: } \tilde{p} \pm \frac{z}{n+z^2}\sqrt{k(n-k)/n + z^2/4}$$

For 95%: acts like adding 2 successes and 2 failures $\Rightarrow \tilde{p} \approx (k+2)/(n+4)$.

Rule of thumb: Use Wald when $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$. Otherwise use Wilson. Many stats packages default to Wilson.

Shortcut (Agresti–Coull): Just use $\tilde{p} = (k+2)/(n+4)$ with a Wald-style CI.

Nearly as good as Wilson, much easier to compute by hand!

CI for Comparing Two Groups

One of the most common tasks: is there a **difference** between two groups?

Two independent samples: $\bar{X}_1$ from group 1 (n_1), $\bar{X}_2$ from group 2 (n_2).

CI for Comparing Two Groups

One of the most common tasks: is there a **difference** between two groups?

Two independent samples: $\bar{X}_1$ from group 1 (n_1), $\bar{X}_2$ from group 2 (n_2).

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

SE of a difference $= \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$ (variances add for independent groups)

CI for Comparing Two Groups

One of the most common tasks: is there a **difference** between two groups?

Two independent samples: $\bar{X}_1$ from group 1 (n_1), $\bar{X}_2$ from group 2 (n_2).

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

SE of a difference $= \sqrt{\mathrm{SE}_1^2 + \mathrm{SE}_2^2}$ (variances add for independent groups)

Example (A/B test): Control ($n_1 = 500$): $\hat{p}_1 = 0.12$. Treatment ($n_2 = 500$): $\hat{p}_2 = 0.15$.

$$\hat{p}_2 - \hat{p}_1 = 0.03, \quad \mathrm{SE} = \sqrt{\frac{0.12 \cdot 0.88}{500} + \frac{0.15 \cdot 0.85}{500}} \approx 0.021 \quad \Rightarrow \quad 95\%\ \text{CI: } [-0.011,\ 0.071]$$

CI for Comparing Two Groups

One of the most common tasks: is there a **difference** between two groups?

Two independent samples: $\bar{X}_1$ from group 1 (n_1), $\bar{X}_2$ from group 2 (n_2).

$$(\bar{X}_1 - \bar{X}_2) \pm z_{\alpha/2} \cdot \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$$

SE of a difference $= \sqrt{\text{SE}_1^2 + \text{SE}_2^2}$ (variances add for independent groups)

Example (A/B test): Control ($n_1 = 500$): $\hat{p}_1 = 0.12$. Treatment ($n_2 = 500$): $\hat{p}_2 = 0.15$.

$$\hat{p}_2 - \hat{p}_1 = 0.03, \quad \text{SE} = \sqrt{\frac{0.12\cdot 0.88}{500} + \frac{0.15\cdot 0.85}{500}} \approx 0.021 \quad \Rightarrow \quad 95\% \text{ CI: } [-0.011,\ 0.071]$$

CI contains 0 $\Rightarrow$ the difference is **not statistically significant** at 95%.
We can't yet conclude the treatment works. (More in Lecture 9.)

Sample Size Planning

Question: How many voters must I survey for a $\pm 3\%$ margin of error at 95% confidence?

Sample Size Planning

Question: How many voters must I survey for a $\pm 3\%$ margin of error at 95% confidence?

Start with the margin of error formula:

$$\text{ME} = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \leq 0.03$$

Sample Size Planning

Question: How many voters must I survey for a $\pm 3\%$ margin of error at 95% confidence?

Start with the margin of error formula:

$$\text{ME} = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \leq 0.03$$

Solve for n (use worst case $p = 0.5$):

$$n \geq \left(\frac{z_{\alpha/2}}{\text{ME}}\right)^2 \cdot p(1-p) = \left(\frac{1.96}{0.03}\right)^2 \cdot 0.25 = 1{,}068$$

Sample Size Planning

Question: How many voters must I survey for a $\pm 3\%$ margin of error at 95% confidence?

Start with the margin of error formula:

$$\text{ME} = z_{\alpha/2} \cdot \sqrt{\frac{p(1-p)}{n}} \le 0.03$$

Solve for n (use worst case $p = 0.5$):

$$n \ge \left(\frac{z_{\alpha/2}}{\text{ME}}\right)^2 \cdot p(1-p) = \left(\frac{1.96}{0.03}\right)^2 \cdot 0.25 = 1{,}068$$

$\text{ME} = \pm 5\%$ $n = 385$	$\text{ME} = \pm 3\%$ $n = 1{,}068$	$\text{ME} = \pm 1\%$ $n = 9{,}604$

The $\sqrt{n}$ law again: $10\times$ more precision requires $100\times$ more data.

The t-Interval and the General MLE Recipe

When σ is unknown (the typical case), use the t-distribution:

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

t_{n-1} has heavier tails than $N(0,1) \Rightarrow$ wider CI. For $n \geq 30$, nearly identical to z.

The t-Interval and the General MLE Recipe

When σ is unknown (the typical case), use the t-distribution:

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

t_{n-1} has heavier tails than $N(0,1) \Rightarrow$ wider CI. For $n \geq 30$, nearly identical to z.

The t-interval is a special case of a **general recipe** that works for any MLE:

General Wald CI for any MLE (from Lecture 7):

$$\hat{\theta}_{\text{MLE}} \quad \pm \quad z_{\alpha/2} \cdot \frac{1}{\sqrt{n \cdot I(\hat{\theta})}}$$

The t-Interval and the General MLE Recipe

When σ is unknown (the typical case), use the t-distribution:

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

t_{n-1} has heavier tails than $N(0,1) \Rightarrow$ wider CI. For $n \geq 30$, nearly identical to z.

The t-interval is a special case of a **general recipe** that works for any MLE:

General Wald CI for any MLE (from Lecture 7):

$$\hat{\theta}_{\text{MLE}} \quad \pm \quad z_{\alpha/2} \cdot \frac{1}{\sqrt{n \cdot I(\hat{\theta})}}$$

Example: Poisson λ. $I(\lambda) = 1/\lambda$, MLE $= \hat{\lambda} = \bar{X}$, SE $= \sqrt{\hat{\lambda}/n}$.

$n = 50$, $\hat{\lambda} = 3.2$: 95% CI $= 3.2 \pm 1.96\sqrt{3.2/50} = [2.70, 3.70]$.

Note: All CIs above are **two-sided**. For one-sided bounds (e.g., "failure rate $\leq U$"), use z_α instead of

Confidence Intervals vs Credible Intervals

Confidence Interval (Frequentist)

"95% of intervals built this way contain θ"

θ is **fixed**; the interval is random.

No prior needed.

$\hat{\theta} \pm z \cdot \mathrm{SE}$

With large n, they often give nearly identical intervals.
The philosophical difference matters most with small n or strong priors.
Neither is "wrong" — they answer **different questions**.

Confidence Intervals vs Credible Intervals

Confidence Interval (Frequentist)

"95% of intervals built this way contain θ"

θ is **fixed**; the interval is random.

No prior needed.

$\hat{\theta} \pm z \cdot \text{SE}$

Credible Interval (Bayesian)

"Given data and prior, θ is in here with 95% prob."

θ is **random**; statement is about θ.

Requires a prior $P(\theta)$.

From the posterior $P(\theta \mid \text{data})$

Confidence Intervals vs Credible Intervals

Confidence Interval (Frequentist)

"95% of intervals built this way contain θ"

θ is **fixed**; the interval is random.

No prior needed.

$\hat{\theta} \pm z \cdot \mathrm{SE}$

Credible Interval (Bayesian)

"Given data and prior, θ is in here with 95% prob."

θ is **random**; statement is about θ.

Requires a prior $P(\theta)$.

From the posterior $P(\theta \mid \text{data})$

With large n, they often give nearly identical intervals.
The philosophical difference matters most with small n or strong priors.
Neither is "wrong" — they answer **different questions**.

The Delta Method: Why Do We Need It?

Situation: We have a CI for p (a proportion), but we *actually* care about the **odds** $\psi = p/(1-p)$.

The Delta Method: Why Do We Need It?

Situation: We have a CI for p (a proportion), but we *actually* care about the **odds** $\psi = p/(1-p)$.

Naive approach: Transform the CI endpoints — just plug them in. But this ignores how the transformation *stretches* the uncertainty. The function g may amplify or compress errors differently at different points.

The Delta Method: Why Do We Need It?

Situation: We have a CI for p (a proportion), but we *actually* care about the **odds** $\psi = p/(1-p)$.

Naive approach: Transform the CI endpoints — just plug them in. But this ignores how the transformation *stretches* the uncertainty. The function g may amplify or compress errors differently at different points.

The delta method gives the principled answer: use a first-order Taylor approximation.

If $\hat{\theta} \dot{\sim} N(\theta, \sigma^2_{\hat{\theta}})$ and g is smooth, then $g(\hat{\theta}) \dot{\sim} N\big(g(\theta),\ [g'(\theta)]^2 \cdot \sigma^2_{\hat{\theta}}\big)$

$$\Rightarrow \quad \textbf{SE of } g(\hat{\theta}) \ \approx\ |g'(\hat{\theta})| \cdot \textbf{SE of } \hat{\theta}$$

The derivative g' tells you how much g "stretches" the uncertainty near $\hat{\theta}$.

Delta Method: Worked Example (Odds)

Setup: $\hat{p} = 0.3$ from $n = 200$. Want a 95% CI for the odds $\psi = p/(1-p)$.

Delta Method: Worked Example (Odds)

Setup: $\hat{p} = 0.3$ from $n = 200$. Want a 95% CI for the odds $\psi = p/(1-p)$.

Step 1: Identify g and compute g'.

$g(p) = p/(1-p), \quad g'(p) = 1/(1-p)^2$

Delta Method: Worked Example (Odds)

Setup: $\hat{p} = 0.3$ from $n = 200$. Want a 95% CI for the odds $\psi = p/(1-p)$.

Step 1: Identify g and compute g'.

$g(p) = p/(1-p), \quad g'(p) = 1/(1-p)^2$

Step 2: Compute the SE of $\hat{p}$ and the SE of $g(\hat{p})$.

$\text{SE}(\hat{p}) = \sqrt{0.3 \times 0.7/200} = 0.0324$

$g'(0.3) = 1/0.7^2 = 2.04 \qquad \Rightarrow \qquad \text{SE}(\hat{\psi}) = 2.04 \times 0.0324 = 0.066$

Delta Method: Worked Example (Odds)

Setup: $\hat{p} = 0.3$ from $n = 200$. Want a 95% CI for the odds $\psi = p/(1-p)$.

Step 1: Identify g and compute g'.

$g(p) = p/(1-p), \quad g'(p) = 1/(1-p)^2$

Step 2: Compute the SE of $\hat{p}$ and the SE of $g(\hat{p})$.

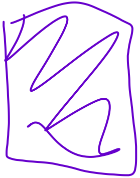
$\text{SE}(\hat{p}) = \sqrt{0.3 \times 0.7/200} = 0.0324$

$g'(0.3) = 1/0.7^2 = 2.04 \quad \Rightarrow \quad \text{SE}(\hat{\psi}) = 2.04 \times 0.0324 = 0.066$

Step 3: Build the Wald CI for ψ.

$\hat{\psi} = 0.3/0.7 = 0.429. \quad$ 95% CI: $0.429 \pm 1.96 \times 0.066 = [0.30,\ 0.56]$

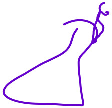
Tip: For asymmetric quantities (odds, hazard ratios), it's often easier to
build a CI on the **log scale**, then exponentiate the endpoints back.



Part II: The Bootstrap

The computational path: let the data speak

Named after the phrase "to pull oneself up by one's bootstraps" —
the data estimates its own sampling distribution, with no outside help.
Introduced by Bradley Efron, Stanford, 1979.



When Formulas Don't Exist

Part I world:

Mean, proportion, MLE
SE formulas exist
Normal approximation works

→

Real world:

Median, correlation, ratio
SE formula is hard or unknown
Distribution may be skewed

→

Bootstrap solution:

Let the *data itself* tell you the SE
No formulas needed

The Core Insight

Ideal world

Draw many samples from the *true* population F

Compute $\hat{\theta}$ on each

$\Rightarrow$ exact sampling distribution

We can't do this

so we do this instead

Bootstrap world

Draw many samples from the *observed data* $\hat{F}_n$

Compute $\hat{\theta}^*$ on each

$\Rightarrow$ approximate sampling dist.

The Core Insight

Ideal world

Draw many samples from the *true* population F

Compute $\hat{\theta}$ on each

$\Rightarrow$ exact sampling distribution

We can't do this → so we do this instead

Bootstrap world

Draw many samples from the *observed data* $\hat{F}_n$

Compute $\hat{\theta}^*$ on each

$\Rightarrow$ approximate sampling dist.

Key idea: The empirical distribution $\hat{F}_n$ is our best estimate of F. Sampling from $\hat{F}_n$ = sampling **with replacement** from the data.

X_1, X_2, X_3

X_3, X_2

X_1, X_1

The Bootstrap Algorithm

Step 1: Start with your original data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

The Bootstrap Algorithm

Step 1: Start with your original data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Step 2: Draw a **bootstrap sample** $\mathbf{x}^*$ by sampling n values **with replacement**.

The Bootstrap Algorithm

Step 1: Start with your original data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Step 2: Draw a **bootstrap sample** $\mathbf{x}^*$ by sampling n values **with replacement**.

Step 3: Compute the statistic on $\mathbf{x}^*$: $\hat{\theta}^* = T(\mathbf{x}^*)$.

The Bootstrap Algorithm

Step 1: Start with your original data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

Step 2: Draw a **bootstrap sample** $\mathbf{x}^*$ by sampling n values **with replacement**.

Step 3: Compute the statistic on $\mathbf{x}^*$: $\hat{\theta}^* = T(\mathbf{x}^*)$.

Step 4: Repeat Steps 2–3 a total of B times (typically $B = 1{,}000$ to 10,000).

The Bootstrap Algorithm

Step 1: Start with your original data $\mathbf{x} = (x_1, x_2, \ldots, x_n)$.

↓

Step 2: Draw a **bootstrap sample** $\mathbf{x}^*$ by sampling n values **with replacement**.

↓

Step 3: Compute the statistic on $\mathbf{x}^*$: $\hat{\theta}^* = T(\mathbf{x}^*)$.

↓

Step 4: Repeat Steps 2–3 a total of B times (typically $B = 1{,}000$ to 10,000).

↓

Step 5: Use the distribution of $\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*$ to estimate SE, bias, or CI.

Visualizing a Bootstrap Sample

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6), \quad n = 7.$

Original:	**2**	**5**	**7**	**3**	**8**	**1**	**6**

Note: some values appear **multiple times**, others **not at all**. That's "with replacement."

Visualizing a Bootstrap Sample

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6)$, $n = 7$.

Original:	**2**	**5**	**7**	**3**	**8**	**1**	**6**	
Boot #1:	**5**	**3**	**3**	**8**	**2**	**5**	**7**	$\hat{\theta}_1^* = \text{median} = 5$

Note: some values appear **multiple times**, others **not at all**. That's “with replacement.”

Visualizing a Bootstrap Sample

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6)$, $n = 7$.

Original:	**2**	**5**	**7**	**3**	**8**	**1**	**6**	
Boot #1:	**5**	**3**	**3**	**8**	**2**	**5**	**7**	$\hat{\theta}^*_1 = \text{median} = 5$
Boot #2:	**7**	**7**	**1**	**6**	**8**	**3**	**8**	$\hat{\theta}^*_2 = \text{median} = 7$

Note: some values appear **multiple times**, others **not at all**. That's "with replacement."

Visualizing a Bootstrap Sample

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6), \quad n = 7.$

Original:	**2**	**5**	**7**	**3**	**8**	**1**	**6**	
Boot #1:	**5**	**3**	**3**	**8**	**2**	**5**	**7**	$\hat{\theta}^*_1 = \text{median} = 5$
Boot #2:	**7**	**7**	**1**	**6**	**8**	**3**	**8**	$\hat{\theta}^*_2 = \text{median} = 7$
Boot #3:	**2**	**1**	**6**	**2**	**3**	**1**	**5**	$\hat{\theta}^*_3 = \text{median} = 2$

Note: some values appear **multiple times**, others **not at all**. That's "with replacement."

Bootstrap SE: A Worked Example

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6)$. Original median $= 5$.

Suppose we run $B = 6$ bootstrap replicates and record the median of each:

Boot medians:	**5**	**7**	**3**	**5**	**6**	**5**

Bootstrap SE: A Worked Example

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6)$. Original median $= 5$.

Suppose we run $B = 6$ bootstrap replicates and record the median of each:

Boot medians:	**5**	**7**	**3**	**5**	**6**	**5**

Step 1: Mean of bootstrap medians:

$$\bar{\theta}^* = \frac{5+7+3+5+6+5}{6} = \frac{31}{6} \approx 5.17$$

Bootstrap SE: A Worked Example

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6)$. Original median $= 5$.

Suppose we run $B = 6$ bootstrap replicates and record the median of each:

Boot medians:	**5**	**7**	**3**	**5**	**6**	**5**

Step 1: Mean of bootstrap medians:

$$\bar{\theta}^* = \frac{5+7+3+5+6+5}{6} = \frac{31}{6} \approx 5.17$$

Step 2: SD of bootstrap medians = bootstrap SE:

$$\widehat{\text{SE}}_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2} = \sqrt{\frac{(5-5.17)^2 + (7-5.17)^2 + \cdots + (5-5.17)^2}{5}} \approx 1.33$$

Bootstrap SE: A Worked Example

Original data: $\mathbf{x} = (2, 5, 7, 3, 8, 1, 6)$. Original median $= 5$.

Suppose we run $B = 6$ bootstrap replicates and record the median of each:

Boot medians:	**5**	**7**	**3**	**5**	**6**	**5**

Step 1: Mean of bootstrap medians:

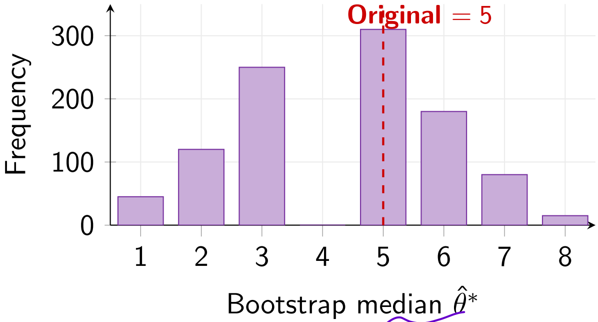
$$\bar{\theta}^* = \frac{5+7+3+5+6+5}{6} = \frac{31}{6} \approx 5.17$$

Step 2: SD of bootstrap medians = bootstrap SE:

$$\widehat{\text{SE}}_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}_b^* - \bar{\theta}^*)^2} = \sqrt{\frac{(5-5.17)^2 + (7-5.17)^2 + \cdots + (5-5.17)^2}{5}} \approx 1.33$$

With $B = 10{,}000$ replicates, we'd get $\widehat{\text{SE}}_{\text{boot}} \approx 1.45$.
Even $B = 6$ gives the right idea!

Bootstrap SE and the Bootstrap Distribution



$$\widehat{\text{SE}}_{\text{boot}} = \text{SD}(\hat{\theta}^*_1, \ldots, \hat{\theta}^*_B)$$

The **spread** of the bootstrap distribution estimates the SE.

No formula needed — just simulation!

How many replicates? $B \geq 200$ for SE, $B \geq 1{,}000$ for percentile CI, $B \geq 5{,}000$ for BCa CI.

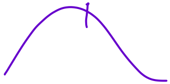
Bootstrap Confidence Intervals: Three Methods

Normal

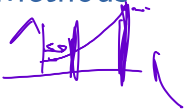
$$\hat{\theta} \pm z_{\alpha/2} \cdot \widehat{\mathsf{SE}}_{\text{boot}}$$

Same as Wald, but with bootstrap SE instead of formula.

Assumes symmetry.



Bootstrap Confidence Intervals: Three Methods



Normal

$$\hat{\theta} \pm z_{\alpha/2} \cdot \widehat{\mathsf{SE}}_{\text{boot}}$$

Same as Wald, but with bootstrap SE instead of formula.

Assumes symmetry.

Percentile

$$[\hat{\theta}^*_{(0.025)}, \quad \hat{\theta}^*_{(0.975)}]$$

Just take the 2.5th and 97.5th percentiles of the bootstrap dist.

Respects asymmetry.

Bootstrap Confidence Intervals: Three Methods

Normal

$$\hat{\theta} \pm z_{\alpha/2} \cdot \widehat{\text{SE}}_{\text{boot}}$$

Same as Wald, but with bootstrap SE instead of formula.

Assumes symmetry.

Percentile

$$[\hat{\theta}^*_{(0.025)}, \quad \hat{\theta}^*_{(0.975)}]$$

Just take the 2.5th and 97.5th percentiles of the bootstrap dist.

Respects asymmetry.

BCa

Adjusted percentiles

Bias-Corrected & Accelerated. Best coverage for complex statistics.

Best general-purpose.

→ increasing sophistication and accuracy

BCa: The Gold Standard Bootstrap CI

BCa = **B**ias-**C**orrected and **a**ccelerated. It adjusts the percentile cutoffs in two ways:

Bias correction (z_0)

If the bootstrap distribution isn't centered on $\hat{\theta}$, the naive percentiles are shifted.

z_0 measures this offset and corrects the cutoff quantiles.

Acceleration (a)

If the SE varies with θ (the sampling distribution changes shape), symmetric cutoffs are wrong.

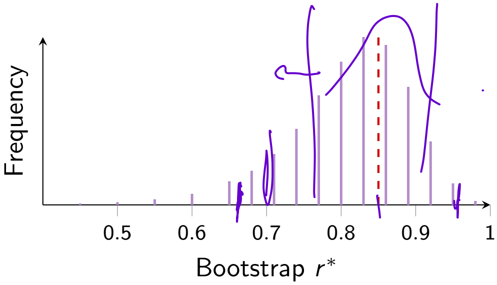
a adjusts for this asymmetry.

Result: better coverage than naive percentile, especially for skewed distributions.

In Python: `scipy.stats.bootstrap()` computes BCa by default.

BCa in Action: A Skewed Example

Setup: Correlation $r = 0.85$ from $n = 15$ pairs. The sampling distribution of r near 1 is **left-skewed**.



Normal: [0.71, 0.99]
Symmetric, overshoots 1!

Percentile: [0.65, 0.94]
Uses raw quantiles.

BCa: [0.68, 0.95]
Corrects for skewness.
Tighter on the right.

BCa adjusts the percentile cutoffs to account for skewness: it shifts the window *toward* the long tail. Result: better coverage in simulation studies.

Why Does the Bootstrap Work?

Real world

$$F \xrightarrow{\text{sample}} \mathbf{x} \xrightarrow{T} \hat{\theta}$$

Variation of $\hat{\theta}$ around θ

$=$ sampling distribution

analogous

Bootstrap world

$$\hat{F}_n \xrightarrow{\text{resample}} \mathbf{x}^* \xrightarrow{T} \hat{\theta}^*$$

Variation of $\hat{\theta}^*$ around $\hat{\theta}$

$=$ bootstrap distribution

Bootstrap principle: The relationship $\hat{F}_n \rightarrow \hat{\theta}^*$ *mimics* the relationship $F \rightarrow \hat{\theta}$.

Formally: $\hat{\theta}^* - \hat{\theta} \approx_d \hat{\theta} - \theta$ as $n \to \infty$. **(Bootstrap consistency.)**

When the Bootstrap Fails

1. Extremes and tails

Max, min, extreme quantiles. Bootstrap can't generate values outside the observed range.

Classic example: $X_i \sim \text{Uniform}(0, \theta)$, $\hat{\theta} = X_{(n)}$. The bootstrap can *never* exceed $X_{(n)}$, so it underestimates variability. (See Lecture 5: MLE for Uniform.)

- **Works well:** Mean, median, correlation, regression coefficients
- **Use with care:** Small n (< 15), mildly heavy tails
- **Avoid:** Extremes (max, min), dependent data, $n < 5$

When the Bootstrap Fails

1. Extremes and tails

Max, min, extreme quantiles. Bootstrap can't generate values outside the observed range.

2. Very small n

With $n = 5$, the "universe" of resamples is too limited to approximate the true distribution.

Classic example: $X_i \sim \text{Uniform}(0, \theta)$, $\hat{\theta} = X_{(n)}$. The bootstrap can *never* exceed $X_{(n)}$, so it underestimates variability. (See Lecture 5: MLE for Uniform.)

- **Works well:** Mean, median, correlation, regression coefficients
- **Use with care:** Small n (< 15), mildly heavy tails
- **Avoid:** Extremes (max, min), dependent data, $n < 5$

When the Bootstrap Fails

1. Extremes and tails

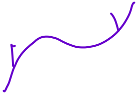
Max, min, extreme quantiles. Bootstrap can't generate values outside the observed range.

2. Very small n

With $n = 5$, the "universe" of resamples is too limited to approximate the true distribution.

3. Dependent data

Time series, spatial data. Naive resampling destroys dependence. Need **block bootstrap**.



Classic example: $X_i \sim \text{Uniform}(0, \theta)$, $\hat{\theta} = X_{(n)}$. The bootstrap can *never* exceed $X_{(n)}$, so it underestimates variability. (See Lecture 5: MLE for Uniform.)

- **Works well:** Mean, median, correlation, regression coefficients
- **Use with care:** Small n (< 15), mildly heavy tails
- **Avoid:** Extremes (max, min), dependent data, $n < 5$

When the Bootstrap Fails

1. Extremes and tails

Max, min, extreme quantiles. Bootstrap can't generate values outside the observed range.

2. Very small n

With $n = 5$, the "universe" of resamples is too limited to approximate the true distribution.

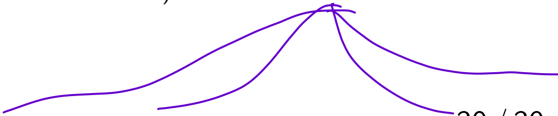
3. Dependent data

Time series, spatial data. Naive resampling destroys dependence. Need **block bootstrap**.

4. Heavy-tailed data

If the population has no finite variance (Cauchy, power-law), bootstrap is **inconsistent**.

Classic example: $X_i \sim \text{Uniform}(0, \theta)$, $\hat{\theta} = X_{(n)}$. The bootstrap can *never* exceed $X_{(n)}$, so it underestimates variability. (See Lecture 5: MLE for Uniform.)



When the Bootstrap Fails

1. Extremes and tails

Max, min, extreme quantiles. Bootstrap can't generate values outside the observed range.

2. Very small n

With $n = 5$, the "universe" of resamples is too limited to approximate the true distribution.

3. Dependent data

Time series, spatial data. Naive resampling destroys dependence. Need **block bootstrap**.

4. Heavy-tailed data

If the population has no finite variance (Cauchy, power-law), bootstrap is **inconsistent**.

Classic example: $X_i \sim \text{Uniform}(0, \theta)$, $\hat{\theta} = X_{(n)}$. The bootstrap can *never* exceed $X_{(n)}$, so it underestimates variability. (See Lecture 5: MLE for Uniform.)

- **Works well:** Mean, median, correlation, regression coefficients
- **Use with care:** Small n (< 15), mildly heavy tails
- **Avoid:** Extremes (max, min), dependent data, $n < 5$

Two Paths, One Goal

When to use which?

Analytical CI vs Bootstrap CI

Analytical (Wald)

$\hat{\theta} \pm z_{\alpha/2} \cdot \text{SE}$

+ Fast, no simulation
+ Exact SE formulas for common cases

– Needs a formula for SE
– Assumes normality, always symmetric

Bootstrap

Resample $\Rightarrow$ percentiles

+ Works for *any* statistic
+ Respects asymmetry (percentile/BCa)

– Computationally expensive
– Fails for extremes, small n

Analytical CI vs Bootstrap CI

Analytical (Wald)

$\hat{\theta} \pm z_{\alpha/2} \cdot \text{SE}$

+ Fast, no simulation
+ Exact SE formulas for common cases

– Needs a formula for SE
– Assumes normality, always symmetric

Bootstrap

Resample ⇒ percentiles

+ Works for *any* statistic
+ Respects asymmetry (percentile/BCa)

– Computationally expensive
– Fails for extremes, small n

In practice: Use analytical CIs when formulas exist and assumptions hold.
Use bootstrap when the statistic is complex (median, correlation, ratio) or skewed.
When both are available, they should agree for large n.

Caution: If you build 20 CIs at 95%, expect ∼1 to miss by chance — **multiple testing correction** needed (Lecture 9).

Bootstrap in Python: It's Simple

```
import numpy as np
from scipy import stats

# Original data
x = np.array([2, 5, 7, 3, 8, 1, 6])
n = len(x)

# Nonparametric bootstrap
B = 10000
boot_medians = np.array([
  np.median(np.random.choice(x, size=n, replace=True))
  for _ in range(B)
])

# Bootstrap SE and percentile CI
se_boot = np.std(boot_medians, ddof=1)
ci_pct = np.percentile(boot_medians, [2.5, 97.5])

# For BCa: scipy.stats.bootstrap() since SciPy 1.7
res = stats.bootstrap((x,), np.median, n_resamples=B)
ci_bca = res.confidence_interval
```

Summary: Quantifying Uncertainty

Confidence interval: A random interval that contains θ with probability $1-\alpha$. It's the *procedure* t

Wald CI: $\hat{\theta} \pm z_{\alpha/2} \cdot \text{SE}$. Uses normal approximation. Need SE formula.

Proportions: Use Wilson (not Wald) for small n or extreme $\hat{p}$.

t-interval: Use when σ is unknown and n is small. Heavier tails $\Rightarrow$ wider CI.

Bootstrap: Resample with replacement $\Rightarrow$ SE and CI without formulas.

Bootstrap CIs: Normal (simple), Percentile (respects shape), BCa (gold standard).

When bootstrap fails: Extremes, tiny n, dependent data, non-smooth statistics.

Two paths: Analytical when formulas exist; bootstrap when they don't. Same goal: quantify unc

Practical: Confidence Intervals & Bootstrap

1. **CI coverage simulation:**
   - Generate $n = 30$ from $N(\mu, \sigma^2)$, compute 95% Wald CI
   - Repeat 10,000 times. What fraction contain μ? (Should be $\approx$ 95%)
   - Try $n = 5$. Still 95%? Now try the t-interval. Better?
2. **Bootstrap vs analytical:**
   - Generate $n = 30$ from Exp(1), compute the median
   - Analytical: no easy formula! Bootstrap: compute $B = 5{,}000$ medians
   - Build percentile and BCa CIs. Does the true median ($\ln 2$) fall in?
3. **Wald vs Wilson for proportions:**
   - Simulate $n = 20$, $p = 0.05$. Compare coverage of Wald and Wilson
   - At what n does Wald catch up to Wilson's coverage?

Homework

1. A factory tests $n = 100$ lightbulbs: $\bar{X} = 1{,}150$ hours, $S = 200$ hours. Construct 90%, 95%, and 99% CIs for the true mean. Which is widest? Why?
2. A poll surveys $n = 500$ voters: 265 support candidate A ($\hat{p} = 0.53$).
   (a) Compute the Wald 95% CI. Can we declare A is leading?
   (b) How many voters do we need for a $\pm 2\%$ margin of error?
3. The heights (cm) of 12 students: 165, 170, 168, 172, 175, 180, 163, 178, 169, 171, 167, 174.
   (a) Compute the bootstrap SE of the **median** using $B = 5{,}000$.
   (b) Construct 95% Normal and Percentile bootstrap CIs.
   (c) Is the bootstrap distribution symmetric? Compare with the analytical CI for the mean.
4. For $X_1, \ldots, X_n \sim \text{Exp}(\lambda)$: MLE $\hat{\lambda} = 1/\bar{X}$, $I(\lambda) = 1/\lambda^2$.
   (a) Build the Wald 95% CI for λ using Fisher information.
   (b) Use the delta method to find a CI for the mean $\mu = 1/\lambda$.
   (c) Compare with a bootstrap percentile CI for λ. Do they agree?

Recommended Visualizations & Resources

Interactive: Confidence Interval Simulation (R Psychologist)

`rpsychologist.com/d3/ci` — watch CIs accumulate in real time. Drag sliders for n, σ, confidence level

Interactive: Seeing Theory — Frequentist Inference (Brown)

`seeing-theory.brown.edu/frequentist-inference` — CI construction and bootstrap resampling animated.

Video: StatQuest — Confidence Intervals & Bootstrapping

Two clear walkthroughs: what CIs are and are not, and how bootstrap builds CIs from data.

Reading: Efron & Tibshirani, "An Introduction to the Bootstrap"

The foundational textbook. Chapters 1–6 cover everything in Part II. Chapters 12–14 cover BCa.

Python: scipy.stats.bootstrap()

BCa intervals in three lines of code (SciPy $\geq$ 1.7). See SciPy docs for details.

Questions?

Next: Lecture 9 — Hypothesis testing, p-values, power, and permutation tests