Stat C131A: Statistical Methods for Data Science

Lecture12: Confidence Intervals

Oct 7 2025

Today

- Wrap up hypothesis testing with completing Confidence Intervals

Confidence Interval

- We know that most of the time, the standardized statistic lies between $-t^*$ and $+t^*$
- $P(-t^* \leq (\bar{X} - \mu)/(s/\sqrt{n}) \leq t^*) = 0.95$
- Here:
  - t^* = critical value from t distribution,
  - Depends on confidence level (e.g., 95%) and $df = n - 1$.

Confidence Interval

$$-t \le \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} \le t$$

$$-t \cdot \frac{s}{\sqrt{n}} \le \overline{X}-\mu \le t \cdot \frac{s}{\sqrt{n}}$$

$$\overline{X} - t \cdot \frac{s}{\sqrt{n}} \le \mu \le \overline{X} + t \cdot \frac{s}{\sqrt{n}}$$

R example for CIs

CI for the difference in the means
(of 2 groups)

- **Central question:** Are two groups truly different on average, or could the observed difference be random noise?
- Examples:
  - Airline A vs Airline B average delay times
  - Fitness app A vs Fitness app B effectiveness (increase in steps walked)
  - …
- Confidence intervals quantify the uncertainty around ($\mu1 - \mu2$)

Why This Matters

- We often compare two populations: schools, drugs, airlines, treatments.
- Observed difference in sample means may not equal true difference.
- CI provides **a plausible range** for the difference in means.
- Not just: 'is there a difference?' but 'how large might it be?'

Hypotheses Context

- **Hypotheses**
  - **Null hypothesis** (**H_0**): $\mu 1 = \mu 2$ (no difference)
  - **Alternative hypothesis** (**H_1**): $\mu 1 \neq \mu 2$ (difference exists)
- CI provides more than yes/no decision, it shows the magnitude
- If 0 is inside CI → no evidence of difference

Mathematically

- $\bar{X} - \bar{Y} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

- $P((\bar{X} - \bar{Y}) - 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X} - \bar{Y}) + 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 0.95$

- $\bar{X} - \bar{Y} \pm 1.96\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Variance

- $$T = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\hat{\sigma}_1^{\,2}}{n_1} + \frac{\hat{\sigma}_2^{\,2}}{n_2}}}.$$
- $$P((\bar{X} - \bar{Y}) - t_{0.975}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{X} - \bar{Y}) - t_{0.025}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}) = 0.95$$

General Formula

- CI formula: $(\bar{x}1 - \bar{x}2) \pm t^* \times SE(\bar{x}1 - \bar{x}2)$
  - $\bar{x}1$ and $\bar{x}2$ are sample means.
  - t^* is the critical value from t distribution
- SE, the Standard Error, is the 'typical random wobble' in $(\bar{x}1 - \bar{x}2)$

Degrees of Freedom

- For pooled test: $df = n1 + n2 - 2$.
- For Welch's test: df ≈ complex adjustment formula.
- df controls cutoff t^* and thus width of CI.

Worked Example: Airlines

- Airline A mean delay = 22 min ($n1 = 30$)
- Airline B mean delay = 14 min ($n2 = 28$)
- Observed difference = 8 min
- Compute SE and CI → example CI = (5, 11)

Interpretation of Example

- 95% CI (5, 11) means true difference plausibly lies between 5.08–11.92 minutes
- Since 0 not included → significant difference
- CI tells **both** significance and effect size

Key Takeaways (Two Means)

- CI gives a range, not just reject/not reject
- If CI includes 0 → no strong evidence of difference.
- Width of CI reflects uncertainty
  - larger n → narrower CI
- Always check assumptions before using formula

BOOTSRAPPING
CONFIDENCE INTERVALS

3.8 Bootstrap Confidence Intervals

- Motivation:
  - What if we were interested in comparing medians ?
    - Why should we compare medians ?
  - What about proportions across 2 groups
    - What is that statistic ?
- Now, formula-based CI **relies on strong assumptions**
  - normality, equal variances, …
- Real data may be skewed, heavy-tailed, or small n.

Bootstrap

- Bootstrap:
  - **resampling** method
  - “assumption-light”
- **Uses the observed sample to approximate the population**

Bootstrapping: the practical need

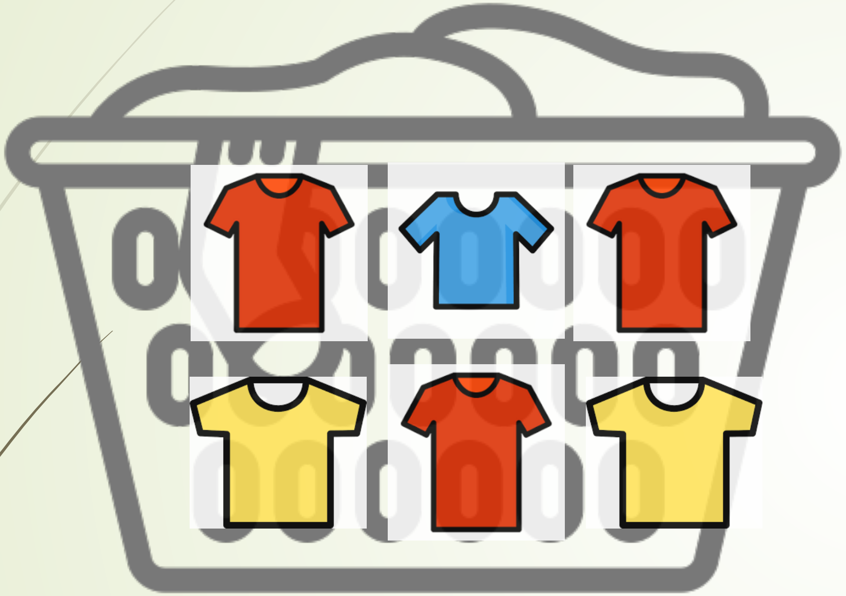
- Hypothesis testing requires knowing the **sampling distribution** of the test statistic to construct confidence intervals or compute p-values
- For means, the **Central Limit Theorem (CLT)** guarantees approximate normality of the difference in means, even for non-normal data
- The CLT's robustness is **unique to the mean** , it does **not** apply directly to medians or most other statistics
- Testing medians, proportions, or other summaries requires **distinct mathematical theory** to derive their null distributions
- Proportions can be treated as means of binary data, allowing modified **t-type parametric tests**.
- Using statistics beyond the mean typically demands **stronger distributional assumptions** and may lack known theory.
- **Bootstrap methods** offer a practical alternative: estimate the statistic's distribution empirically via resampling, bypassing analytical derivations.
- Thus, bootstrapping generalizes inference to settings where **theoretical distributions are unknown or intractable**

Core Idea of Bootstrapping

- Treat your sample as a stand-in for the population
- Draw many bootstrap samples **with replacement**
- Each bootstrap sample **has same size as original**
- Compute statistic of interest for each resample.

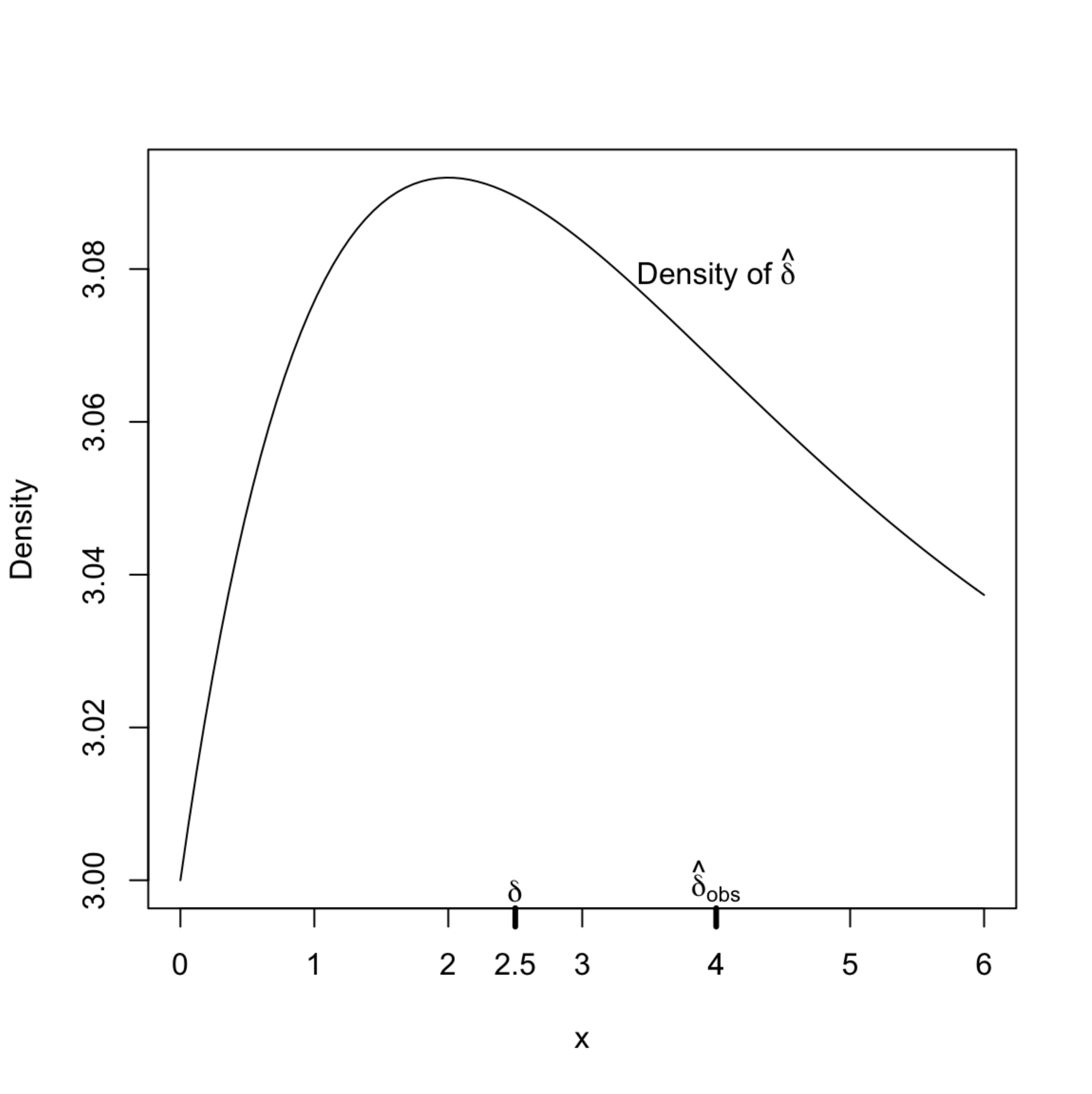
Step-by-Step Bootstrap Procedure

- 1. Start with observed dataset.
- 2. Resample with replacement, size = n.
- 3. Compute statistic (e.g., $\bar{x}1 - \bar{x}2$).
- 4. Repeat thousands of times (1000–10,000).
- 5. Use bootstrap distribution to form CI.



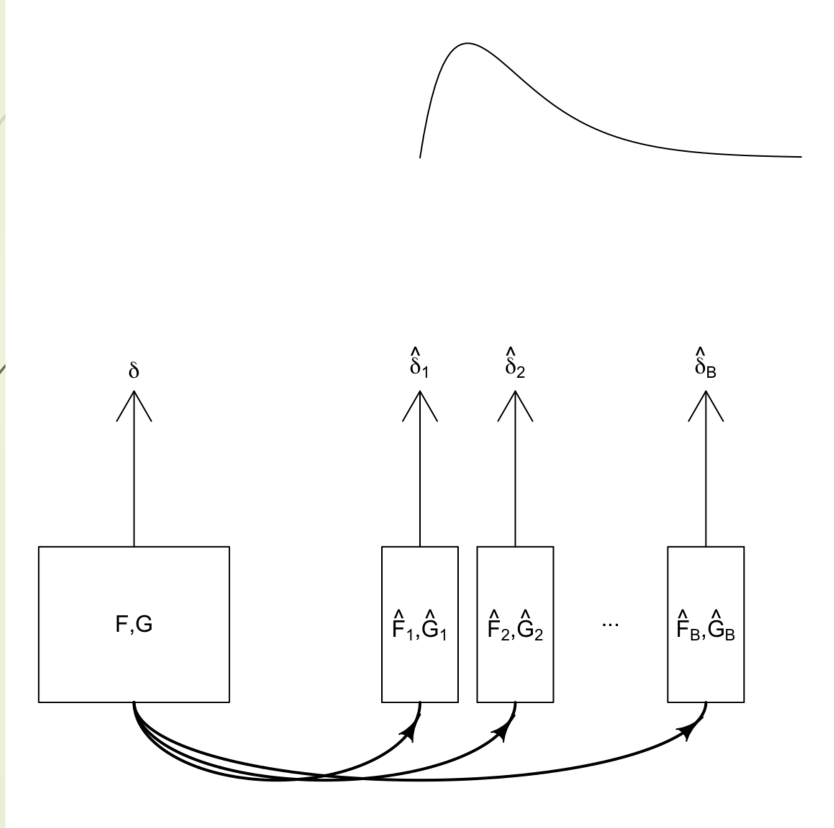
Types of Bootstrap CIs

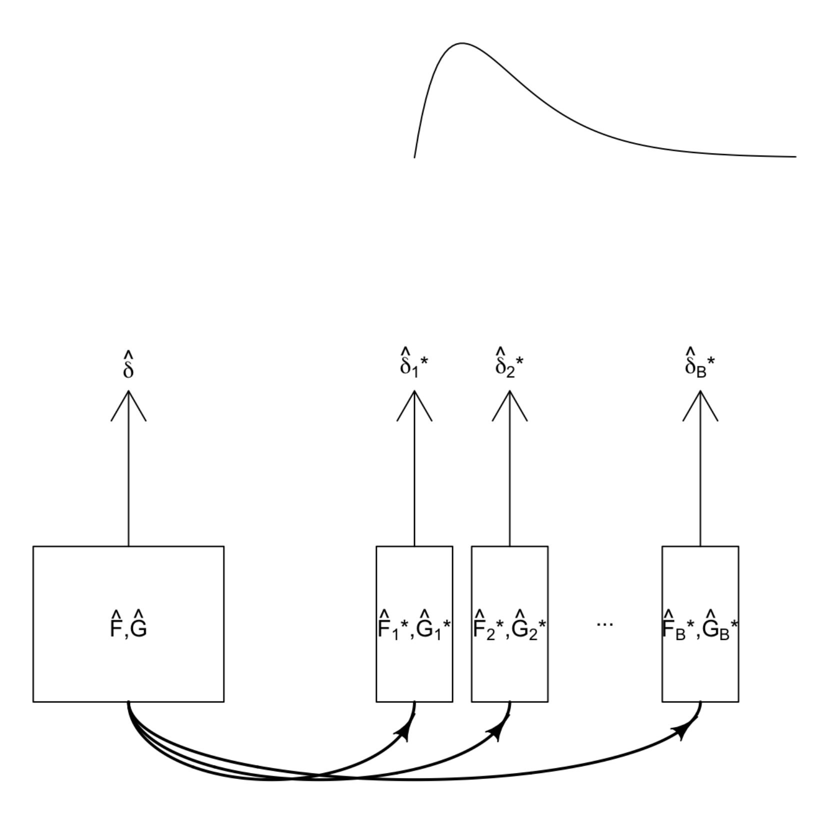
- Percentile method: middle 95% of bootstrap statistics.
- SE method: estimate bootstrap SE, use ± z* cutoff.
- Bias-Corrected and Accelerated (BCa): adjusts for skew and bias.
- $0.95 = P(\hat{\delta} - w_1 \leq \delta \leq \hat{\delta} + w_2)$
- $(V_1 = \hat{\delta} - w_1, V_2 = \hat{\delta} + w_2)$
- $P(\delta - w_2 \leq \hat{\delta} \leq \delta + w_1)$

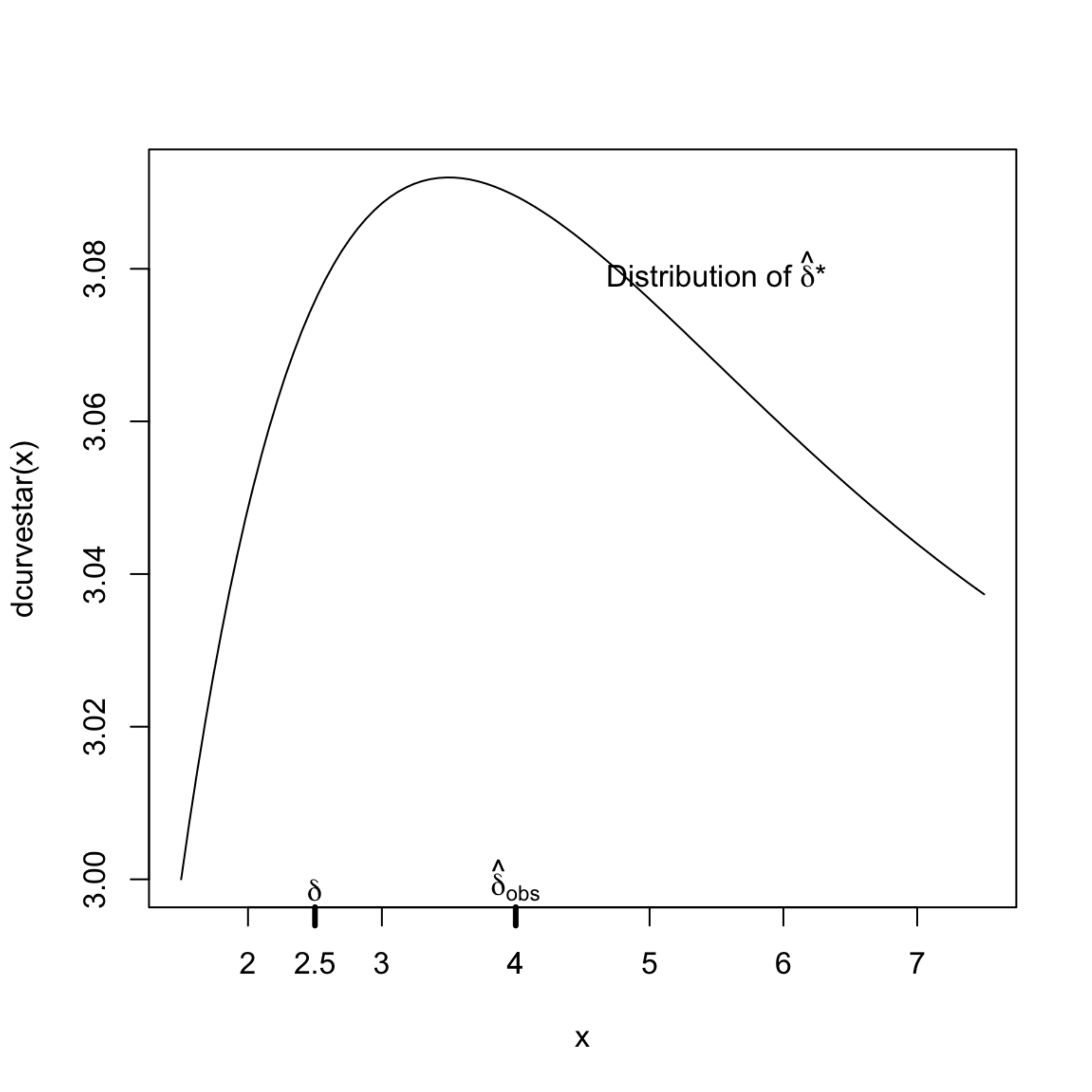


Density of $\hat{\delta}$
Density
3.00
3.02
3.04
3.06
3.08
δ
$\hat{\delta}_{obs}$
0
1
2
2.5
3
4
5
6
x

Bootstrap: difference of means







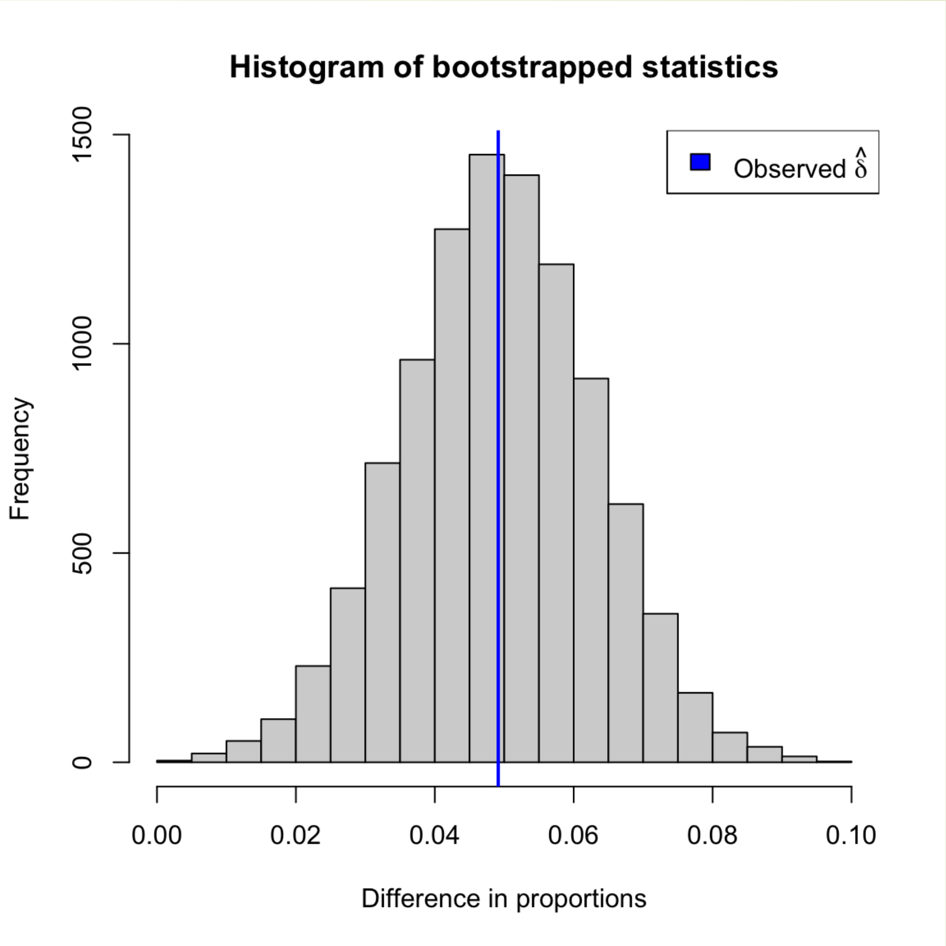
Distribution of $\hat{\delta}^*$
dcurvestar(x)
3.00
3.02
3.04
3.06
3.08
δ
$\hat{\delta}_{obs}$
2
2.5
3
4
5
6
7
x

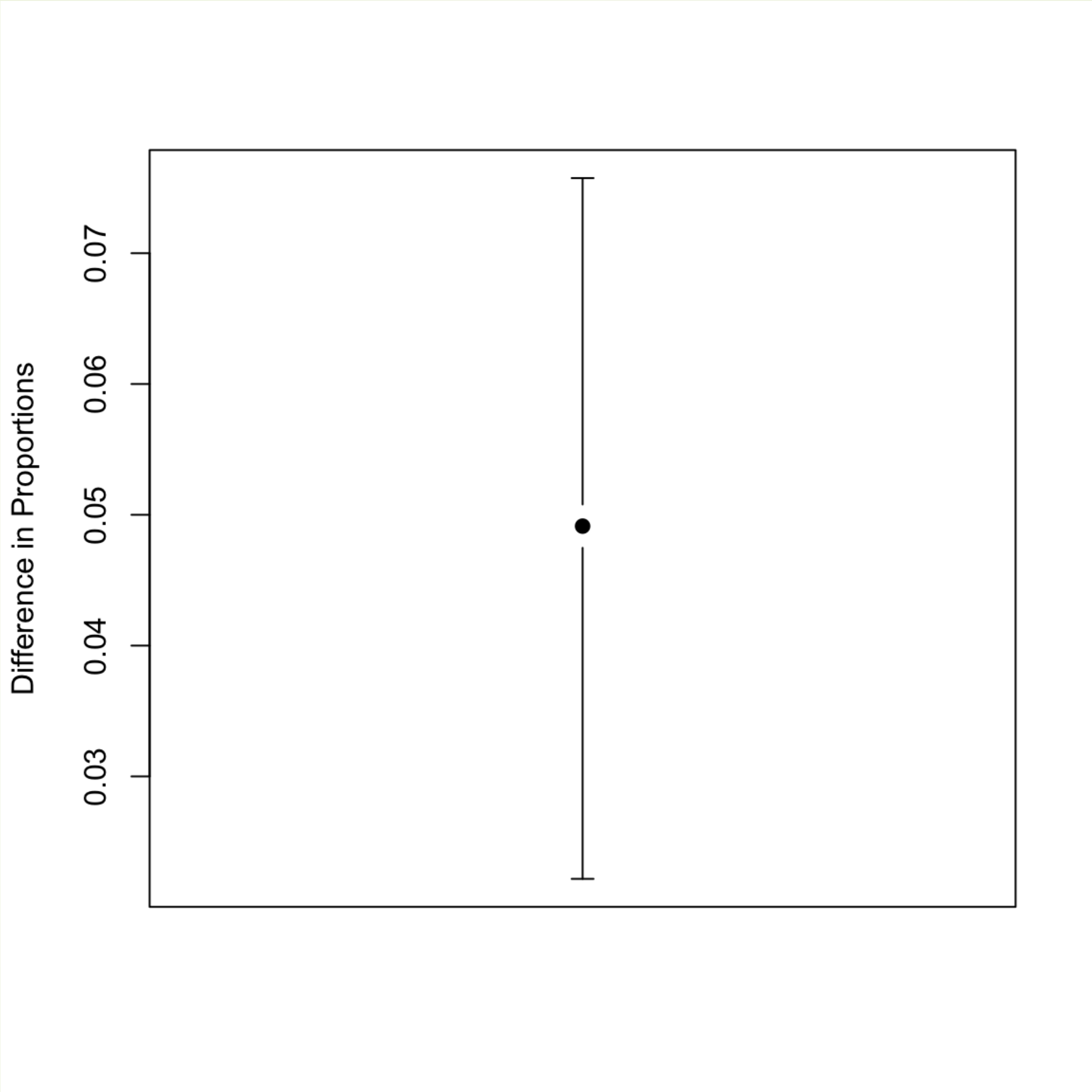
Mathematically

- $P(|\hat{\delta} - \delta| > 1) \approx P(|\hat{\delta}^* - \hat{\delta}| > 1)$
- $0.95 = P(\hat{\delta} - W_2^* \leq \hat{\delta}^* \leq \hat{\delta} + W_1^*)$
- $0.95 \approx P(\delta - W_2^* \leq \hat{\delta} \leq \delta + W_1^*) = P(\hat{\delta} - W_1^* \leq \delta \leq \hat{\delta} + W_2^*)$
- $(\hat{\delta} - W_1^*, \hat{\delta} + W_2^*)$

Bootstrap Example: Airlines

- Resample Airline A delays 1000 times, Airline B delays 1000 times.
- Each time compute mean difference ($\bar{x}1 - \bar{x}2$).
- Form distribution of bootstrap differences.
- **95% CI = central 95% of bootstrap distribution**.




Difference in Proportions
0.03
0.04
0.05
0.06
0.07

Comparing Bootstrap-based CI with Formula-based CI

- Formula CI: fast, assumption-heavy.
- Bootstrap CI: slower, assumption-light, more flexible.
- Large n, normal data → similar results.
- Small n or skewed data → bootstrap more reliable.

Why Bootstrap Works

- Bootstrap mimics repeated sampling.
- Empirical distribution approximates true sampling distribution.
- Assumes observed sample is representative.
- Leverages the data you already have to quantify uncertainty.

Key Takeaways (Bootstrap)

- CI for two means estimates how different groups are.
- Bootstrap extends CI to messy/non-normal data.
- Both approaches quantify uncertainty.
- Always report both estimate and CI → communicates effect size + precision.

Specific Statistics, and Distributions (or lack thereof)

Statistic	Known Distribution	Type	Method of Derivation
Mean	Normal (exact/asymptotic)	CLT	Analytical
Difference of means	t / Normal	CLT + estimated variance	Analytical
Variance	χ^2	Exact under Normality	Analytical
Ratio of variances	F	Derived from χ^2	Analytical
Proportion	Binomial / Normal	CLT	Analytical
Correlation	t / Normal (Fisher z)	Transform theory	Analytical
Rank-based tests	Normal (large n)	Asymptotic	Approximate
Median / Percentiles	No simple form	Depends on F	Requires bootstrap

Bootstrapping: Perspective

- Bootstrapping estimates a statistic's sampling distribution by **resampling with replacement**, it works numerically for any statistic, linear or not.
- A **closed-form (mathematical)** solution for variance or confidence intervals exists **only when the statistic is a smooth / differentiable functional** of the population distribution.
- **Linear or differentiable statistics** (for example, the mean or regression coefficients):
  - The statistic changes smoothly with small perturbations in the data.
- **Non-smooth statistics** (for example, the median or quantiles):
  - Bootstrap works only if the density at that point is positive and continuous
    - $f(m) > 0$
    - If $f(m) = 0$ or the distribution is discrete or flat, the bootstrap becomes inconsistent.
- **Highly non-smooth or non-local functionals** (for example, the mode):
  - The mapping $F \to \text{mode}(F)$ is discontinuous.
  - The ordinary bootstrap fails and requires smoothed or specialized variants.
- Bootstrapping is always computable, but only smooth, differentiable statistics have a valid mathematical limit law