Lecture 14: Bootstrap Methods

Module 4: part 2

Spring 2025

Logistics

- Module 3 Assessment due Mar 16

Why Bootstrap? A Nutritional Science Perspective

The Challenge in Nutritional Research:

- You are studying the relationship between dietary polyphenol intake and inflammation markers.
- Small sample size: Only 28 participants.
- Highly variable responses between individuals.
- Non-normal distribution of inflammation markers (right-skewed).
- Presence of influential observations.

Traditional Approaches Fall Short:

- Parametric tests require normality assumptions.
- Transformations distort interpretability.
- Small sample prevents reliable asymptotic approximations.
- Impossible to collect more data (budget constraints).

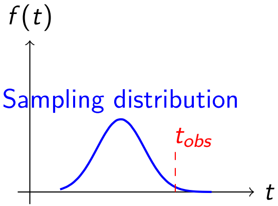
Sampling Distribution?

How to estimate confidence intervals for the effect size when assumptions are violated?

Model Assumption Violations

Model-Based Hypothesis Testing

- Hypothesis testing compares our test statistic to a **hypothetical sampling distribution**.
- **Sampling Distribution:** Distribution of the test statistic if we repeated data collection infinitely.
- **Model-based approach:** Data generating assumptions determine the theoretical sampling distribution.
- **Robust approaches:** Methods that work even when assumptions are violated.



Key Challenge

What happens when our model assumptions are violated?

- **Sandwich Estimator:** Robust standard error estimation that allows for heteroskedasticity.
- Also known as **Huber-White** or **heteroskedasticity-consistent (HC)** standard errors.

Monte Carlo Methods: Introduction

- Computational technique to study properties of random processes
- **Historical Context:**
  - Developed during Manhattan Project (1940s)
  - Named after Monaco's Monte Carlo casino district
  - Pioneered by Stanislaw Ulam and John von Neumann
- Used to calculate statistical properties: mean, variance, quantiles, distribution shapes

Basic Strategy

1. Simulate data generation process many times.
2. Calculate desired statistics from samples.
3. Increase simulation count for higher precision.

Example: Blackjack Strategy

Simulate thousands of blackjack games → Calculate win percentage → Estimate long-term success rate.

Monte Carlo Methods: Applications in Statistical Inference

Advantages:

- Verifies theoretical properties
- Handles complex models
- Provides visual understanding
- Tests robustness to violations

Limitations:

- Requires known data generation process
- Dependent on strong assumptions
- Computationally intensive
- May not reflect real-world complexity

Key Questions

- Can we "approximately" draw new data without exact models?
- Can we relax assumptions while maintaining validity?

Bridge to Bootstrapping

This leads us to resampling methods like bootstrapping, which use observed data to approximate sampling distributions without strong parametric assumptions.

Bootstrap

Bootstrap Procedures: Approximating Sampling Distributions

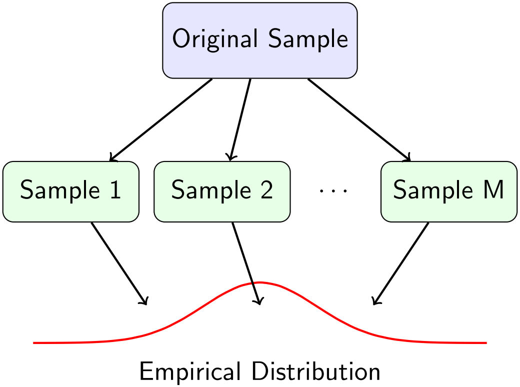
Definition: Bootstrap methods approximate the sampling distribution of a statistic by resampling from observed data.

- Requires **weaker assumptions** than parametric methods
- Particularly valuable for **small sample sizes**
- Handles **non-standard statistics** where theoretical distributions are unknown
- Provides **empirical confidence intervals** without normality assumptions

Key Insight

Bootstrap treats the sample as a "mini-population" that approximates the true population.

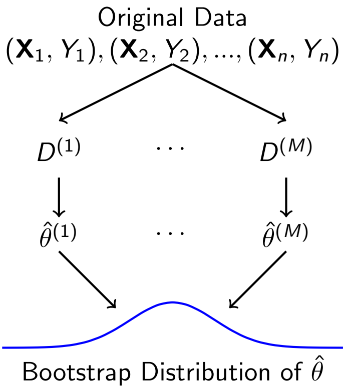
Bootstrap plot



Empirical Bootstrap: Step-by-Step Implementation

Given: Data pairs $(\mathbf{X}_i, Y_i)$ for $i = 1, \ldots, n$

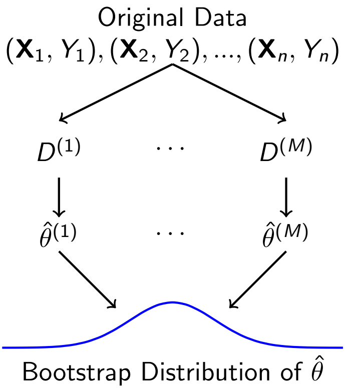
1. Calculate statistic $\hat{\theta}$ from original data
2. Resample n observations **with replacement** from original data $\rightarrow$ create $D^{(m)}$
3. Calculate $\hat{\theta}^{(m)}$ from bootstrapped dataset $D^{(m)}$
4. Repeat steps 2-3 for $m = 1, \ldots, M$ (typically $M \geq 1000$)



Empirical Bootstrap: Step-by-Step Implementation

Given: Data pairs $(\mathbf{X}_i, Y_i)$ for $i = 1, \ldots, n$

1. Calculate statistic $\hat{\theta}$ from original data
2. Resample n observations **with replacement** from original data $\rightarrow$ create $D^{(m)}$
3. Calculate $\hat{\theta}^{(m)}$ from bootstrapped dataset $D^{(m)}$
4. Repeat steps 2-3 for $m = 1, \ldots, M$ (typically $M \geq 1000$)

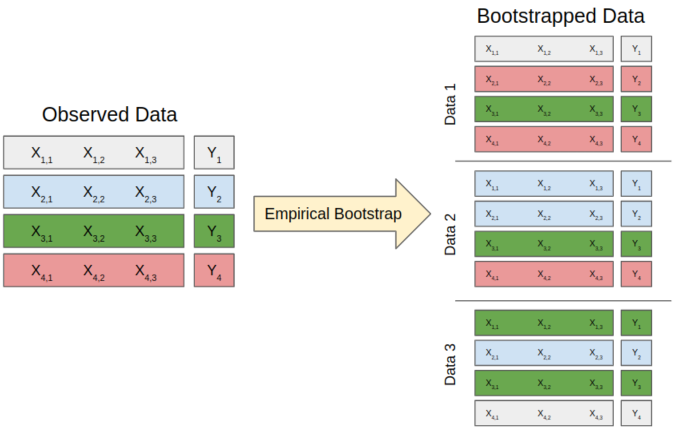


Why It Works

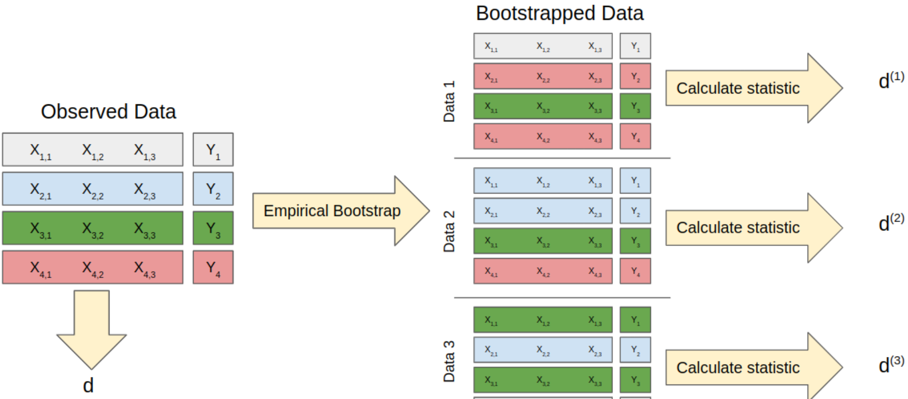
Sampling with replacement mimics the original data generating process, approximating the true sampling distribution.

Empirical/Case/Pairs Bootstrap

Empirical/Case/Pairs Bootstrap



Empirical/Case/Pairs Bootstrap



Wild Bootstrap: Handling Heteroskedasticity

1. Calculate statistic d from observed data
2. Fit regression model: $\hat{y}_i = \hat{b}_0 + \sum_k \hat{b}_k x_{i,k}$
3. Compute residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
4. Create bootstrap samples $D^{(m)}$ with:

$$y_i^{(m)} = \hat{b}_0 + \sum_k \hat{b}_k x_{i,k} + \hat{\varepsilon}_i \times Z_i$$

   where $Z_i \sim N(0, 1)$ or alternative distribution
5. Calculate test statistic $d^{(m)}$ from each bootstrap sample
6. Repeat steps 4-5 for $m = 1, \ldots, M$ samples

Why It Works

Using fitted values maintains the regression structure, while multiplying residuals by random noise preserves the heteroskedastic error pattern at each point.

Wild Bootstrap: Handling Heteroskedasticity

1. Calculate statistic d from observed data
2. Fit regression model: $\hat{y}_i = \hat{b}_0 + \sum_k \hat{b}_k x_{i,k}$
3. Compute residuals: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
4. Create bootstrap samples $D^{(m)}$ with:

$$y_i^{(m)} = \hat{b}_0 + \sum_k \hat{b}_k x_{i,k} + \hat{\varepsilon}_i \times Z_i$$

   where $Z_i \sim N(0, 1)$ or alternative distribution
5. Calculate test statistic $d^{(m)}$ from each bootstrap sample
6. Repeat steps 4-5 for $m = 1, \ldots, M$ samples

Why It Works

Using fitted values maintains the regression structure, while multiplying residuals by random noise preserves the heteroskedastic error pattern at each point.

Key Advantage

Preserves heteroskedasticity pattern in the original data!

Bootstrap Methods: Practical Concerns

Sample Size and Computation

- M should be large (≥ 1000).
- Larger M reduces MC error.
- Parallel computing can help.

Design Considerations

- For fixed X, empirical bootstrap performs poorly.
- With outliers, use robust bootstrap variants.

Method Selection Guidelines

Issue	Recommended Method
Heteroskedasticity	Wild bootstrap
Fixed X	Wild bootstrap
Non-linear relationship	Model-based bootstrap
Time series	Block bootstrap
Outliers	Robust bootstrap
Clustered data	Cluster bootstrap

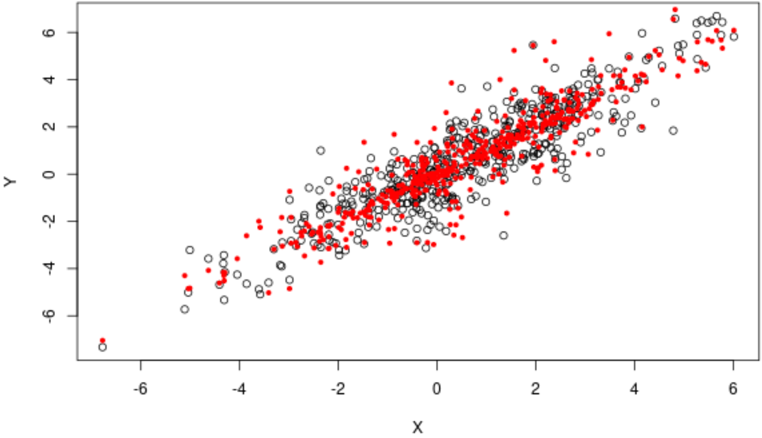
Diagnostics

- Check bootstrap distribution shape.
- Compare different bootstrap methods.
- Assess sensitivity to M.

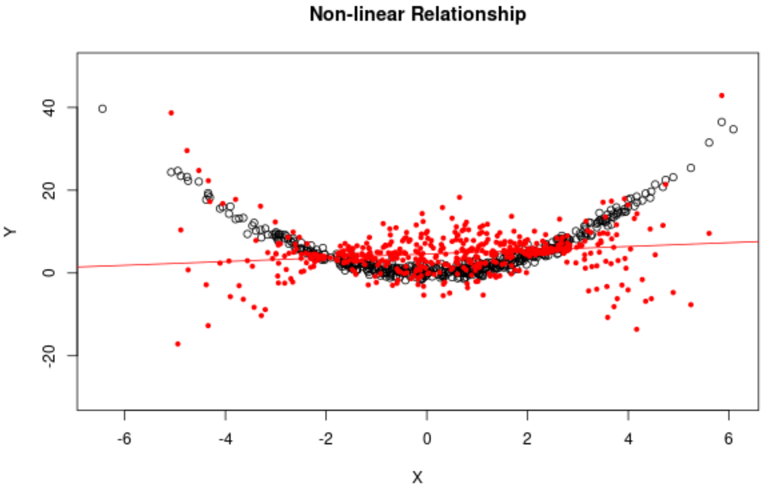
Limitations of Wild Bootstrap

- Assumes correct model specification
- May not work well with highly asymmetric error distributions
- Performance depends on multiplier distribution choice
- Less effective for very small samples ($n < 20$)

Practical Concerns



Practical Concerns



Bootstrap Variance Estimation & Inference

Given M bootstrap samples with test statistics $\{d^{(1)}, d^{(2)}, \ldots, d^{(M)}\}$:

1. Calculate bootstrap mean:
$$\bar{d}^\star = \frac{1}{M}\sum_{m=1}^{M} d^{(m)}$$
2. Estimate variance:
$$\widehat{\text{Var}}(\hat{d}) = \frac{1}{M}\sum_{m=1}^{M}(d^{(m)} - \bar{d}^\star)^2$$
3. Construct confidence interval:
$$\hat{b}_1 \pm t_{n-p-1,1-\alpha/2}\sqrt{\widehat{\text{Var}}(\hat{b}_1)}$$
4. Test hypothesis $H_0 : b_1 = \beta$:
$$t = \frac{\hat{b}_1 - \beta}{\sqrt{\widehat{\text{Var}}(\hat{b}_1)}}$$

   Compare to t_{n-p-1} distribution for p-value.

Alternative: Percentile Method

Directly use empirical quantiles of bootstrap distribution:

$$\text{CI}_{1-\alpha} = \left[d^{(\lfloor \alpha/2 \cdot M \rfloor)}, d^{(\lceil (1-\alpha/2)\cdot M \rceil)}\right]$$

where $d^{(k)}$ is the k-th ordered bootstrap statistic

Bootstrap Variance Estimation & Inference

Important Considerations

- For skewed distributions, percentile method may be preferred.
- For highly non-normal statistics, transformations before bootstrap may improve performance.
- Bootstrap variance estimate converges to true variance as $M \to \infty$ and $n \to \infty$.

Percentile CI

Alternatively, we can use a "percentile approach" to form a $1-\alpha$ confidence interval

1. Calculate the statistic $\hat{d}$ from the observed data
2. Let $\delta^{(m)} = d^{(m)} - \hat{d}$
3. Let $\delta_{\alpha/2}$ be the $\alpha/2$ quantile of $\delta^{(m)}$ for $m = 1, \ldots M$
4. Let $\delta_{1-\alpha/2}$ be the $1-\alpha/2$ quantile of $\delta^{(m)}$ for $m = 1, \ldots M$
5. Construct the confidence interval as

$$(\hat{d} - \delta_{1-\alpha/2}, \hat{d} - \delta_{\alpha/2})$$

Non-standard quantities

The bootstrap also allows us to compute confidence intervals for “non-standard” quantities

- Bootstrap can be used for a wide variety of statistics (i.e., d and $d^{(m)}$ can be many different quantities of interest)
- If you can compute the quantity of interest from data, the bootstrap distribution (can in most cases) be used to approximate the sampling distribution

Non-standard quantities

The bootstrap also allows us to compute confidence intervals for "non-standard" quantities

- Bootstrap can be used for a wide variety of statistics (i.e., d and $d^{(m)}$ can be many different quantities of interest)
- If you can compute the quantity of interest from data, the bootstrap distribution (can in most cases) be used to approximate the sampling distribution

Example:

- Suppose I'm interested in the quantity b_1/b_2
- The sampling distribution of $\hat{b}_1/\hat{b}_2$ is hard to describe theoretically
- Use bootstrap to approximate sampling distribution of $\hat{b}_1/\hat{b}_2$

Simulation study

$$Y_i = b_1 X_{i,1} + b_2 X_{i,2} + \varepsilon_i$$

$$(X_{i,1}, X_{i,2}) \sim \text{Correlated Gamma} \qquad \varepsilon_i \mid X_i = N(0, x_{i,1}^2)$$

- Create a 95% confidence interval for b_1 using
  - model based standard errors
  - wild bootstrap (percentile method)
  - empirical bootstrap (percentile method)
  - empirical bootstrap (Bootstrapped variance estimate)
- Create a 95% confidence interval for b_1/b_2 using empirical bootstrap (percentile)

Simulation study

Since this is a simulation and we know the truth, we can measure the proportion of times the 95% confidence interval actually covers the parameter of interest

n	MB	WB (P)	EB (P)	EB (V)	b_1/b_2
50	0.63	0.81	0.89	0.87	0.91
100	0.58	0.86	0.90	0.90	0.91
250	0.58	0.91	0.93	0.92	0.92
500	0.57	0.93	0.93	0.94	0.93

Bootstrap

- Bootstrap is a powerful concept which allows us to approximate the sampling distribution
- Can be used under weaker assumptions
- Lots of research on how to improve bootstrap and adapt to different settings
- Can be used for non-standard quantities