# Week 5 Notes: The Bootstrap

Antonio R. Linero
University of Texas at Austin

# Goals

1. Learn how to use the bootstrap to estimate the sampling distribution of (almost) any statistic $T = T(\mathcal{D})$.
2. Learn the distinction between various types of bootstraps (nonparametric and parametric bootstraps).
3. Get practice implementing the bootstrap and apply it to several problems.

# Motivation

**Goal:** estimate the *sampling distribution* of a statistic

$$T = T(\mathcal{D})$$

Why? If $T$ is a "good" estimator of a parameter $\psi$ then

$$T \pm z_{\alpha/2} \times \text{std. error}(T)$$

will be a "good" confidence interval for $\psi$.

# Motivation

**Goal:** estimate the *sampling distribution* of a statistic

$$T = T(\mathcal{D})$$

Why? If $T$ is a "good" estimator of a parameter $\psi$ then

$$T \pm z_{\alpha/2} \times \text{std. error}(T)$$

will be a "good" confidence interval for $\psi$.

Other types of intervals are of course possible, and bootstrap can help with those as well.

# Motivation: Robust Standard Errors

Assume we fit a GLM such that

$$g(\mu_i) = X_i^\top \beta$$

*correctly describes the mean model.* How does MLE behave?

# Motivation: Robust Standard Errors

Assume we fit a GLM such that

$$g(\mu_i) = X_i^\top \beta$$

*correctly describes the mean model.* How does MLE behave?

1. $\sqrt{N}(\widehat{\beta} - \beta_0) \to N(0, \Sigma)$ ($\sqrt{N}$-consistent!)
2. $\Sigma \neq$ the inverse Fisher information.

# Motivation: Robust Standard Errors

Problem: given $X_1, \dots, X_N \overset{\text{iid}}{\sim} F$, what if I want a CI for

$$m_F = \text{median}(F)?$$

Several ways to do this! But one is to use the fact that

$$M \overset{\bullet}{\sim} \text{Normal}\{m_F, \sigma_M^2\} \quad \text{where} \quad M = \text{median}(X_1, \dots, X_N).$$

But hard to estimate $\sigma_M^2$!

# Motivation: Robust Standard Errors

Problem: given $X_1, \ldots, X_N \stackrel{\text{iid}}{\sim} F$, what if I want a CI for

$$m_F = \text{median}(F)?$$

Several ways to do this! But one is to use the fact that

$$M \stackrel{\bullet}{\sim} \text{Normal}\{m_F, \sigma^2_M\} \quad \text{where} \quad M = \text{median}(X_1, \ldots, X_N).$$

But hard to estimate $\sigma^2_M$!

More generally: many estimators are *asymptotically linear*, meaning that

$$\widehat{\theta} = \theta_0 + \frac{1}{N} \sum_i \phi(X_i; \theta_0) + o_P(N^{-1/2})$$

for a mean-0 function $\phi(X_i; \theta_0)$ called the *influence function.* Implies a CLT, but may be difficult to estimate variance!

# The Bootstrap Principle

**The Bootstrap Principle:**

Suppose that $\mathcal{D} \sim G$, $\psi = \psi(G)$ is a parameter of $G$ we are interested in, and $T = T(\mathcal{D})$ is an estimator of $\psi$. Then we can approximate the sampling distribution of $T(\mathcal{D})$ by

1. estimating $G$ with some $\widehat{G}$; and

2. using the sampling distribution of $T^\star = T(\mathcal{D}^\star)$ to estimate the sampling distribution of $T$, where $\mathcal{D}^\star \sim \widehat{G}$.

# Example

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$.

- Estimate $\mu \approx \bar{X}$ and $\sigma^2 \approx s^2 = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$.

# Example

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$.

- Estimate $\mu \approx \bar{X}$ and $\sigma^2 \approx s^2 = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$.
- $\widehat{F} = \text{Normal}(\bar{X}, s^2)$.

# Example

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$.

- Estimate $\mu \approx \bar{X}$ and $\sigma^2 \approx s^2 = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$.
- $\widehat{F} = \text{Normal}(\bar{X}, s^2)$.
- $\mathcal{D}^\star = \{X_1^\star, \ldots, X_N^\star\}$ with $X_i^\star \overset{\text{iid}}{\sim} \text{Normal}(\bar{X}, s^2)$.

# Example

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$.

- Estimate $\mu \approx \bar{X}$ and $\sigma^2 \approx s^2 = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$.
- $\widehat{F} = \text{Normal}(\bar{X}, s^2)$.
- $\mathcal{D}^\star = \{X_1^\star, \ldots, X_N^\star\}$ with $X_i^\star \overset{\text{iid}}{\sim} \text{Normal}(\bar{X}, s^2)$.
- From properties of normal $\bar{X}^\star \sim \text{Normal}(\bar{X}, s^2/N)$.

# Example

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} \text{Normal}(\mu, \sigma^2)$.

- Estimate $\mu \approx \bar{X}$ and $\sigma^2 \approx s^2 = \frac{\sum_i (X_i - \bar{X})^2}{N-1}$.
- $\widehat{F} = \text{Normal}(\bar{X}, s^2)$.
- $\mathcal{D}^\star = \{X_1^\star, \ldots, X_N^\star\}$ with $X_i^\star \overset{\text{iid}}{\sim} \text{Normal}(\bar{X}, s^2)$.
- From properties of normal $\bar{X}^\star \sim \text{Normal}(\bar{X}, s^2/N)$.
- Approximate sampling variance of $\bar{X}$ is therefore $s^2/N$.

# Example: Nonparametric Bootstrap

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$. What is the distribution of $M = \text{median}(X_1, \ldots, X_N)$?

- Estimate $F$ with $\widehat{F} = \frac{1}{N} \sum_i \delta_{X_i}$.

# Example: Nonparametric Bootstrap

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$. What is the distribution of $M = \text{median}(X_1, \ldots, X_N)$?

- Estimate $F$ with $\widehat{F} = \frac{1}{N} \sum_i \delta_{X_i}$.
- Approximate sampling distribution of $M$ with $\text{median}(X_1^\star, \ldots, X_N^\star)$ where $X_i^\star \overset{\text{iid}}{\sim} \widehat{F}$.

# Example: Nonparametric Bootstrap

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$. What is the distribution of $M = \text{median}(X_1, \ldots, X_N)$?

- Estimate $F$ with $\widehat{F} = \frac{1}{N} \sum_i \delta_{X_i}$.
- Approximate sampling distribution of $M$ with $\text{median}(X_1^\star, \ldots, X_N^\star)$ where $X_i^\star \overset{\text{iid}}{\sim} \widehat{F}$.
- Cannot do this in closed form…

# Example: Nonparametric Bootstrap

Suppose $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$. What is the distribution of $M = \text{median}(X_1, \ldots, X_N)$?

- Estimate $F$ with $\widehat{F} = \frac{1}{N} \sum_i \delta_{X_i}$.
- Approximate sampling distribution of $M$ with $\text{median}(X_1^\star, \ldots, X_N^\star)$ where $X_i^\star \overset{\text{iid}}{\sim} \widehat{F}$.
- Cannot do this in closed form…
- So use Monte Carlo instead! Sample many new datasets, compute the median on each, and use the resulting empirical distribution of $M$!

# Bootstrap Variance Estimation

The bootstrap can be used to approximate the variance (or standard deviation) of $T$ as follows:

1. Draw $\mathcal{D}^\star \sim \widehat{G}$ (for example, if $X_1, \ldots, X_N \overset{\text{iid}}{\sim} F$ then we take $X_1^\star, \ldots, X_N^\star \overset{\text{iid}}{\sim} \mathbb{F}$).
2. Compute $T^\star = T(\mathcal{D}^\star)$.
3. Repeat steps 1 and 2 $B$ times to get $T_1^\star, \ldots, T_B^\star$.
4. Let $v_{\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} \left( T_b^\star - \bar{T} \right)^2$ where $\bar{T} = \frac{1}{B} \sum_b T_b^\star$.

# Pseudocode

```r
## Let X be a vector of size N, sampled from F
T <- median(X)
Tboot <- numeric(N)
for(i in 1:B) {
  ## Sample N iid draws from the empirical distribution of X
  Xstar <- sample(x = X, size = N, replace = TRUE)
  ## Save the result
  Tboot[i] <- median(Xstar)
}
v_boot <- var(Tboot)
se_boot <- sqrt(v_boot)
```

# The Normal Interval

For asympototically linear statistics, it makes sense to use the interval

$$T \pm z_{\alpha/2} \times \mathrm{se}_{\mathrm{boot}}$$

Great if we are close to normality! Can also use a $t_{\alpha/2}$ interval if worried about Monte Carlo error...

# Basic Percentile Method

Mimics Bayesian reasoning:

1. Sample many $T^\star$'s according to the nonparametric bootstrap.
2. Use the interval $(T^\star_{\alpha/2}, T^\star_{1-\alpha/2})$ where $T^\star_\gamma$ denotes the $100\gamma^{\text{th}}$ percentile of $T^\star$.

Why does this work?

# Pivotal Intervals

- Don't approximate the sampling distribution of $T$ with that of $T^\star$
- Instead, approximate the sampling distribution of $\zeta = T - \psi$ with $\zeta^\star = T^\star - T$.
- Leads to the interval

$$\psi \in (T - \zeta^\star_{1-\alpha/2}, T - \zeta^\star_{\alpha/2}) = (2T - T^\star_{1-\alpha/2}, 2T - T^\star_{\alpha/2}).$$

Show on board why this works.

# Pivotal Intervals

- Don't approximate the sampling distribution of $T$ with that of $T^\star$
- Instead, approximate the sampling distribution of $\zeta = T - \psi$ with $\zeta^\star = T^\star - T$.
- Leads to the interval

$$\psi \in (T - \zeta^\star_{1-\alpha/2}, T - \zeta^\star_{\alpha/2}) = (2T - T^\star_{1-\alpha/2}, 2T - T^\star_{\alpha/2}).$$

Show on board why this works.

*Works best when $T - \psi$ is a* ***pivotal quantity***. Exact pivots are hard to get without strong assumptions, however…

# Exercise

**Exercise: Wasserman 8.1**

Consider the following dataset:

```
df <- data.frame(
  LSAT = c(576, 635, 558, 578, 666, 580, 555, 661, 651,
           605, 653, 575, 545, 572, 594),
  GPA = c(3.39, 3.30, 2.81, 3.03, 3.44, 3.07, 3.00, 3.43,
          3.36, 3.13, 3.12, 2.74, 2.76, 2.88, 3.96)
)
```

which are LSAT scores (for entrance to law school) and GPA. Estimate the standard error of the correlation coefficient $\rho$ using the bootstrap. Find a 95 percent confidence interval using the normal, pivotal, and percentile methods.

# Exercise

**Exercise: Wasserman 8.2**

Conduct a simulation to compare the various bootstrap confidence interval methods. Let $N = 50$ and let $\psi = \frac{1}{\sigma^3}\int (x-\mu)^3 F(dx)$ be the skewness. Draw $Y_1, \ldots, Y_N \sim \text{Normal}(0,1)$ and set $X_i = e^{Y_i}$, $i = 1, \ldots, N$. Construct the three types of bootstrap 95 percent intervals for $\psi$ from the data $X_1, \ldots, X_N$. Repeat this whole thing many times and estimate the true coverage of the three intervals.