Survey of Advanced Bootstrap Topics

BIOS 6618

CU Anschutz

Bootstrap Review

Refresher: Two-Sample Bootstrap (Case Resampling)

Bootstrap sampling *mimics how the data were obtained*. For an experiment designed to compare two populations, we randomly take a sample *from each*. Hence, the bootstrap sample will mimic this process:

Given independent samples of sizes m and n from two populations,

1. Draw a resample of size m with replacement from the first sample and a separate resample of size n with replacement from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Bootstrap Review

With our case resampling bootstraps, we reviewed:

- One- and two-sample problems
- Calculating the standard error for a statistic
- Calculating the bootstrap percentile confidence interval
- Calculating the normal percentile confidence interval

However, there were potential limitations and the bootstrap world is much, much larger. In this lecture we introduce some advanced topics.

Alternative Confidence Interval Calculations

Overview

The bootstrap percentile confidence interval is easy to implement (e.g. take the $\alpha/2$ and $1-\alpha/2$ percentiles of your bootstrap distribution), but may be inaccurate (e.g., using the $|\text{bias}|/\text{SE} < 0.1$ rule of thumb).

The normal percentile confidence interval estimates the standard error from our bootstrap distribution, but ultimately assumes normality and may be inaccurate (e.g., coverage in each tail departing from the desired $\alpha/2$ rate).

In contrast, there are various modifications to bootstrap confidence intervals that have been proposed to further improve our estimation.

Bias-Corrected (BC) Intervals

The general idea behind the BC interval is to adjust our bootstrap percentile CI to account for bias.

More technically, it is built on the idea that the estimator $\hat{\theta}$ may not be median-unbiased (which is similar to the concept of mean-unbiasedness where $E(\hat{\theta}) = \theta$ but for the median).

The deviations from median-unbiasedness can be estimated in our bootstrap since we assume $\hat{\theta}$ is the true value.[1]

Efron proposed a median-bias corrected interval that allows for a shift in the distribution of $g(\hat{\theta})$ by an unknown amount z_0, which in our bootstrap setting is:

$$P^*\{g(\hat{\theta^*}) - g(\hat{\theta}) + z_0 \leq x\} \approx \Phi(x)$$

[1] Boos, D. D., & Stefanski, L. A. (2013). *Essential statistical inference: theory and methods*, Ch. 11.6. New York: Springer.

Bias-Corrected Accelerated (BC_a) Intervals

A second improvement proposed by Efron is the BC_a interval. It attempts to account for underlying higher order effects so that it corrects our bootstrap percentile CI for both bias and skew.

It adds an "acceleration" constant a that is related to the 3rd moment skewness coefficient:

$$P^* \left\{ \frac{g(\hat{\theta^*}) - g(\hat{\theta})}{1 + ag(\hat{\theta})} + z_0 \leq x \right\} \approx \Phi(x)$$

Studentized t-Intervals

The studentized t (or bootstrap t) interval is another alternative approach to estimating confidence intervals. It estimates quantiles from the bootstrap distribution of the Student's t-test:

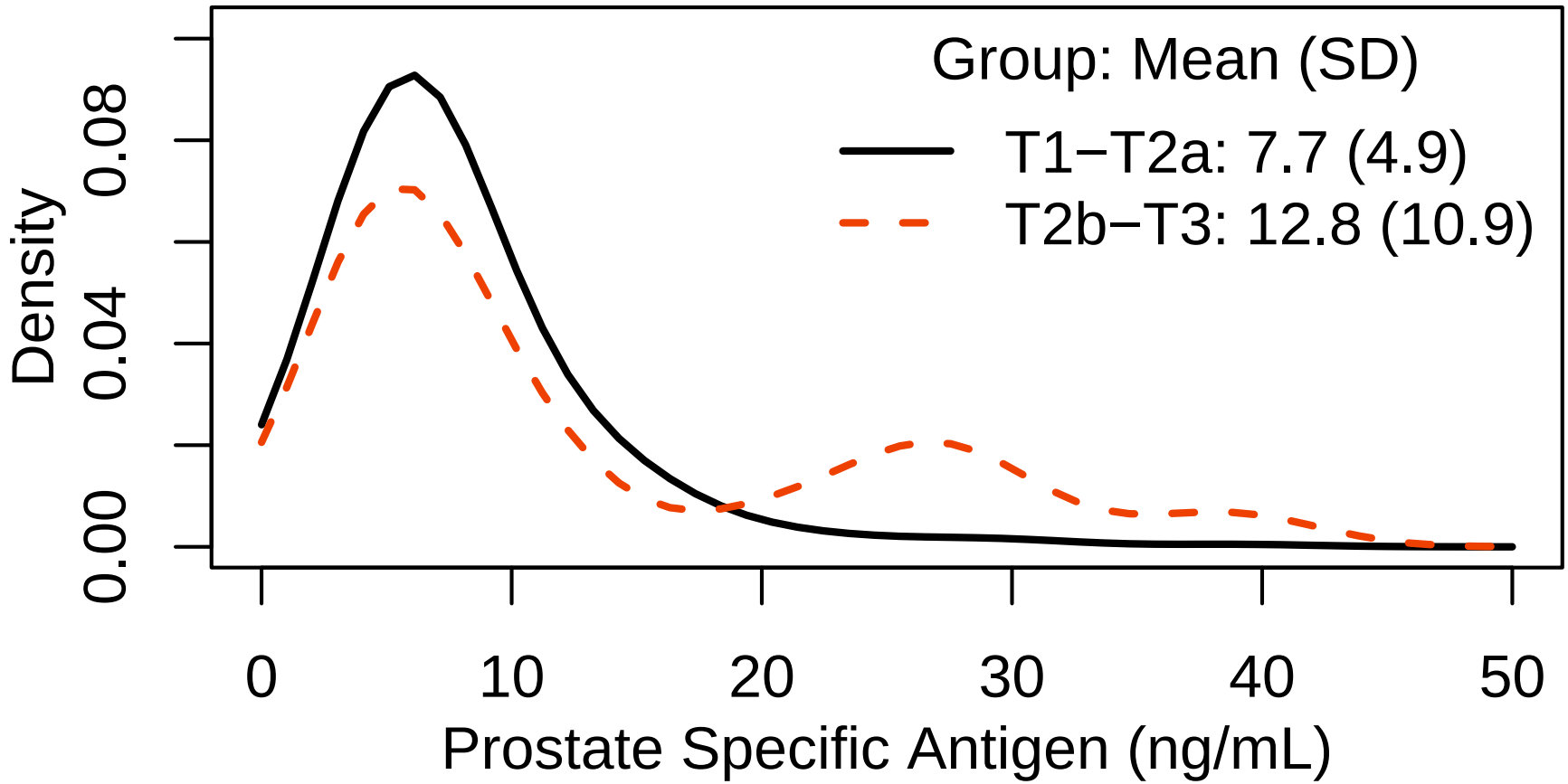
$$(\hat{\theta} - t^*_{(1-\alpha/2)}\hat{\text{se}}_\theta, \hat{\theta} - t^*_{(\alpha/2)}\hat{\text{se}}_\theta)$$

$\hat{\theta}$ is the estimate from our sample, $t^*_{(1-\alpha/2)}$ is the $1-\alpha/2$ percentile based on the bootstrap Student's t-test (i.e., $t^* = (\hat{\theta}^* - \hat{\theta})/\hat{\text{se}}_{\hat{\theta}^*}$), $\hat{\text{se}}_\theta$ is the SE from the sample, and $\hat{\text{se}}_{\hat{\theta}^*}$ is the SE from the bootstrap iteration.

The resulting intervals do not have to be symmetric since $t^*_{(1-\alpha/2)} \neq -t^*_{(\alpha/2)}$, which is an improvement over the normal percentile CI. However, it can be influenced by outliers, in which case the percentile intervals may be more appropriate.

Bootstrap CI Example in R

From the TSHS[2] Blood Storage dataset, we will explore if there is a mean difference in preoperative prostate specific antigen (PSA; ng/mL) between tumor stages (T1-T2a [X, m=267] versus T2b-T3 [Y, n=34]):



[2] https://www.causeweb.org/tshs/category/dataset/

Bootstrap CI Example in R

```
library(boot)
set.seed(6618)
mean_diff <- function(dat, index){
  m1 = mean(subset(dat[index, 1], dat[index, 2] == 1)) #T1-T2a
  m2 = mean(subset(dat[index, 1], dat[index, 2] == 2)) #T2b-T3
  return(m1 - m2)}
boot_res <- boot(data = dat[,c('PreopPSA','T.Stage')],
                 statistic = mean_diff, R=10000)
boot.ci( boot_res )
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = boot_res)
##
## Intervals :
## Level      Normal              Basic
## 95%   (-8.885, -1.397 )   (-8.712, -1.258 )
##
## Level     Percentile            BCa
## 95%   (-9.015, -1.561 )   (-9.566, -1.895 )
## Calculations and Intervals on Original Scale
```

Other Types of Bootstraps

Overview

We have focused on the nonparametric case-resampling bootstrap approach, which is fairly straightforward and works for many problems.

However, there are many other approaches to bootstrapping. In this section we briefly touch on some you may encounter and general details on the motivation and approach.

Parametric Bootstrap

A parametric bootstrap assumes our data comes from a known distribution with unknown parameters. One can estimate the unknown parameters from the available data, then use simulation to generate new samples.

This approach works well if we have a strong reason to believe the data follows a particular distribution, resulting in more accurate estimates of the standard error and CI. This is especially true for smaller sample sizes.

However, if our assumed distribution is wrong it may lead to less accurate estimates.

Parametric Bootstrap

The general steps of a parametric bootstrap are as follows:

1. Assume our data $X_1, ..., X_m$ are drawn from a parametric distribution $F(\theta)$
2. Estimate θ by a statistic $\hat{\theta}$ (e.g., μ from $\bar{X}$)
3. Generate B bootstrap samples by simulating from $F(\hat{\theta})$ (e.g., simulate m cases from $N(\bar{X}, s^2)$)
4. Calculate θ^* for each bootstrap sample

We can then use the B estimates of θ^* to estimate the SE or CI, just like the nonparametric bootstrap.

Smoothed Bootstrap

Nonparametric bootstraps do not work as well for rank-based estimators, such as the median.

One potential solution is to add a small amount of zero-centered random noise to each resampling observation in each of the B bootstrap samples.

Commonly used distributions are Uniform$(-\delta, \delta)$ or $N(0, \delta)$, where δ may depend on the context. Too small and not enough smoothing occurs, but too large and it obscures the original information.

The results can be summarized to estimate the SE or CI, just like the nonparametric bootstrap.

Residual Bootstraps

In linear regression, we may wish to treat our X's as "fixed" rather than random (as is assumed with case resampling) based on our study design.

One easy way to address this is to resample based on residuals using the following steps:

1. Fit the model and keep $\hat{Y}_i$ and $\hat{e}_i = Y_i - \hat{Y}_i$.
2. For each observation set $(Y_i, X_{i1}, \cdots, X_{im})$, add a randomly resampled with replacement residual to the fitted value: $Y_i^* = \hat{Y}_i + \hat{e}_j^*$.
3. Refit your regression model with outcomes of Y_i^* and save the statistic(s) of interest.
4. Repeat B times.

The results can be summarized to estimate the SE or CI, just like the nonparametric bootstrap.

Wild Bootstraps

Related to the residual bootstrap approach, but the wild bootstrap relaxes the assumption of homoscedasticity by multiplying the randomly sampled residuals by another random variable, v_i with mean 0 and variance 1:

1. Fit the model and keep $\hat{Y}_i$ and $\hat{e}_i = Y_i - \hat{Y}_i$.
2. For each observation set $(Y_i, X_{i1}, \cdots, X_{im})$, add a randomly resampled with replacement residual to the fitted value: $Y_i^* = \hat{Y}_i + v_i \hat{e}_j^*$.
3. Refit your regression model with outcomes of Y_i^* and save the statistic(s) of interest.
4. Repeat B times.

Common choices of v_i include $N(0,1)$ or *Mammen's two-point distribution* which is $-(\sqrt{5}-1)/2$ with probability $(\sqrt{5}+1)/(2\sqrt{5})$ and $(\sqrt{5}+1)/2$ with probability $(\sqrt{5}-1)/(2\sqrt{5})$.

The results can be summarized to estimate the SE or CI, just like the nonparametric bootstrap.

Advanced Bootstrap Summary

In this lecture we examined alternative approaches to calculating confidence intervals adjusting for bias (BC) or bias and skewness (BC_a).

We also briefly reviewed some alternative bootstrap strategies that may be used in certain contexts.

There are many more bootstrap approaches and confidence interval strategies that may be helpful in different settings, so keep your eyes out for others.