STAT 641: BOOTSTRAPPING METHODS

Jiyoun Myung

Department of Statistics and Biostatistics
California State University, East Bay

Spring 2021, Day 6

Chapter 10: Estimates of Bias

Estimates of bias

We have concentrated on standard error as a measure of accuracy for an estimator $\hat{\theta}$. There are other useful measures of statistical accuracy (or statistical error), measuring different aspects of $\hat{\theta}$'s behavior.

This chapter concerns bias, the difference between the expectation of an estimator $\hat{\theta}$ and the quantity θ being estimated.

Real world vs Bootstrap world

An unknown probability distribution F has given data $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ by random sampling, $F \to \mathbf{x}$.

- The bias of $\hat{\theta} = s(\mathbf{x})$ as an estimate of θ defined to be the difference between the expectation of $\hat{\theta}$ and the value of the parameter θ,

$$\text{bias}_F = \text{bias}_F(\hat{\theta}, \theta) = E_F[s(\mathbf{x})] - t(F).$$

- The bootstrap estimate of bias is defined to be the estimate $\text{bias}_{\hat{F}}$,

$$\text{bias}_{\hat{F}} = E_{\hat{F}}[s(\mathbf{x}^*)] - t(\hat{F}).$$

Here $t(\hat{F})$, the plug-in estimate of θ, may differ from $\hat{\theta} = s(\mathbf{x})$.

Example

- If $s(\mathbf{x})$ is the mean and $t(F)$ is the population mean, then $\text{bias}_F = 0$. and also $\text{bias}_{\hat{F}} = 0$.

- If $s(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, then $\text{bias}_F = -\frac{1}{n}\sigma_F^2$ and

  $\text{bias}_{\hat{F}} = -\frac{1}{n^2}\sum_{i=1}^{n}(x_i - \bar{x})^2$.

The bootstrap estimate of bias

For most statistics that arise in practice, the ideal bootstrap estimate $\text{bias}_{\hat{F}}$ must be approximated by Monte Carlo simulation.

We generate independent bootstrap samples $\mathbf{x}^{*1}, \mathbf{x}^{*2}, \cdots, \mathbf{x}^{*B}$, evaluate the bootstrap replications $\hat{\theta}^*(b) = s(\mathbf{x}^{*b})$, and approximate the bootstrap expectation $E_{\hat{F}}[s(\mathbf{x}^*)]$ by the average

$$\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B = \sum_{b=1}^{B} s(\mathbf{x}^{*b})/B.$$

The **bootstrap estimate of bias** based on the B replications $\widehat{\text{bias}}_B$ is

$$\widehat{\text{bias}}_B = \hat{\theta}^*(\cdot) - t(\hat{F}).$$

We can calculate both $\widehat{\text{se}}_B$ and $\widehat{\text{bias}}_B$ from the same set of bootstrap replications.

Example

Sample Variance

```
set.seed(123)
n <- 25 # small sample size
x <-  rnorm(n) # random sample from the standard normal
varx <- var(x) * (n-1)/n # Biased estimator for variance

# sample variance, bias_F, biis_{empirical F}
c(varx, -1/n, -varx/n)
```

```
## [1]  0.86045016 -0.04000000 -0.03441801
```

```
B <- 1000 # Number of required bootstrap samples
boot_var <- rep(0, B)
for(i in 1: B){
xstar <- sample(x, n, replace = TRUE)
boot_var[i] <- var(xstar)*(n-1)/n
}

# resample variance and bias estimate
c(mean(boot_var), mean(boot_var) - varx)
```

```
## [1]  0.82874637 -0.03170378
```

Bias Correction

Why would we want to estimate the bias of $\hat{\theta}$?

The usual reason is to correct $\hat{\theta}$ so that it becomes less biased. If $\widehat{\text{bias}}$ is an estimate of $\text{bias}_F(\hat{\theta}, \theta)$, then the obvious **bias-corrected estimator** is

$$\bar{\theta} = \hat{\theta} - \widehat{\text{bias}}.$$

Taking $\widehat{\text{bias}}$ equal to $\widehat{\text{bias}}_B = \hat{\theta}^*(\cdot) - \hat{\theta}$ gives

$$\bar{\theta} = \hat{\theta} - \widehat{\text{bias}}_B = 2\hat{\theta} - \hat{\theta}^*(\cdot).$$

- Bias estimation is usually interesting and worthwhile, but the exact use of a bias estimate is often problematic. Authors recommend using:

a. $\hat{\theta}$ if $\widehat{\text{bias}}_B$ is small relative to the $\widehat{\text{se}}_B$.

b. $\bar{\theta}$ if $\widehat{\text{bias}}_B$ is large relative to the $\widehat{\text{se}}_B$.

c. They suggest and justify that if $\widehat{\text{bias}}_B < .25\widehat{\text{se}}_B$ then the bias can be ignored (unless the goal is precise confidence interval estimation using this standard error).

Example : Law data

```r
library(bootstrap) # law data
library(boot)
theta.hat <- function(d,i) cor(d[i, 1], d[i, 2])

#Perform bootstrapping using the boot function.
set.seed(123)
boot_corr <- boot(data = law, statistic = theta.hat , R = 5000)
boot_corr
```
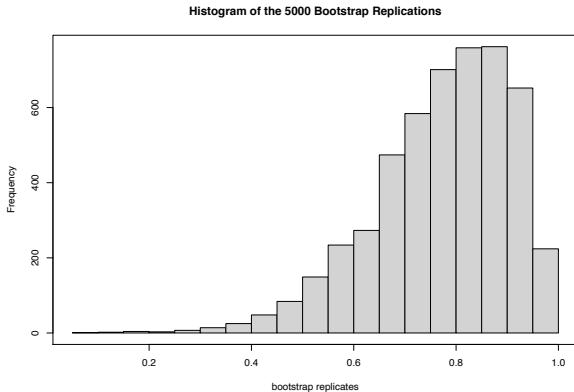
```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = law, statistic = theta.hat, R = 5000)
##
##
## Bootstrap Statistics :
##      original      bias    std. error
## t1* 0.7763745 -0.003927276   0.1332268
```

```r
(mean(boot_corr$t) - boot_corr$t0 )/sd(boot_corr$t)
```

```
## [1] -0.02947813
```

Example

Distribution of bootstrap replicates

Histogram of the 5000 Bootstrap Replications

The bootstrap distribution seems shifted to the right.

Example

```
#The estimated bias
mean(boot_corr$t)-boot_corr$t0
```

```
## [1] -0.003927276
```
```
#The estimated se
sd(boot_corr$t)
```

```
## [1] 0.1332268
```
```
# true correlation coefficient
cor(law82[,2], law82[,3])
```

```
## [1] 0.7599979
```
```
#sample correlation coefficient
boot_corr$t0
```

```
## [1] 0.7763745
```
```
#The bias corrected estimate
boot_corr$t0-(mean(boot_corr$t)-boot_corr$t0)
```

```
## [1] 0.7803018
```

- The biased corrected estimate is further from the true theta than the uncorrected estimate, so correcting for the bias was harmful in this case.

- Bias correction can be dangerous in practice, due to high variability in $\widehat{\text{bias}}$.

Bootstrap Confidence Intervals

Percentile Method

Some Background

For many years, statisticians could not set confidence intervals on many parameters of interest without having to make strong and often unrealistic assumptions about the distribution from which the data were obtained.

- There is theory that tells how one can set a confidence interval on the sd, provided the data come from a normal distribution. But if one is interested on the sd of income in the U.S., we know from the histogram that there is a very long right tail. Income is not normally distributed, but economists still need to estimate the sd.

- There is theory on how to estimate confidence intervals for the ratio of two expected values, provided that both the numerator and the denominator are from independent normal random variables. But for many applications, this is untrue—income per hour worked is an example.

Standard confidence interval

Given an estimate $\hat{\theta}$ and an estimated standard error $\hat{\text{se}}$, the usual 95% confidence interval for θ is

$$\hat{\theta} \pm 1.96 \cdot \hat{\text{se}}.$$

For example, suppose we have data $x_1, x_2, \ldots, x_n$. If we know the data was drawn from $N(\mu, \sigma^2)$ with the unknown mean μ and known variance σ^2 then we have seen that

$$\left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}},\ \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

is a 95% confidence interval for μ.

Bootstrap Confidence Intervals

Percentile Method

- The interval between the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution of a statistic is a 95% bootstrap percentile confidence interval for the corresponding parameter.
- If the bias of the bootstrap distribution is small and the distribution is close to Normal, the bootstrap t and percentile confidence intervals will agree closely.
- If they do not agree, this is evidence that the Normality and bias conditions are not met. Neither type of interval should be used if this is the case.

Comparing Two Samples

Mouse Data: a two-sample problem, not a paired sample problem.

Group			Survival time (in days)						Mean
Treatment	94	197	16	38	99	141	23		86.86
Control	52	104	146	10	51	30	40	27	46	56.22

- The data seem to support that the treatment may have a beneficial effect on survival. To assess this more formally, we need to consider whether the two populations of survival times could have the same mean.

Comparing Two Samples

Let F and G denote the population distributions of the treatment and no treatment data respectively.

The statistical question of interest is whether the means of F and G are equal, i.e. whether $\theta(F, G) = \mu(F) - \mu(G) = 0$ in the obvious notation. As usual for the nonparametric bootstrap, we aim to answer this question using the information contained only in the observed data, i.e. the samples of sizes $n_F = 7$ and $n_G = 9$ from F and G respectively.

Let $x_1, x_2, \ldots, x_7$ represent the treatment data, and $y_1, y_2, \ldots, y_9$ denote the no-treatment data. It is reasonable to assume that the x_i values are an independent sample from F, that the y_j values are an independent sample from G, and that the x_i and y_j values are mutually independent. From the observed data, we have a single value of the random variable

$$\hat{\theta}_{7,9}(F, G) = \bar{x}_7 - \bar{y}_9 = 86.86 - 56.22 = 30.64,$$

which is the difference between the sample means of independent samples of sizes 7 and 9 from F and G respectively.

Comparing Two Samples

Construct the bootstrap distribution

1. Draw a resample of size $n_F = 7$ with replacement form the first sample and a separate resample of size $n_G = 9$ from the second sample. Compute the statistic that compares the two groups, such as the difference between the two sample means.

2. Repeat the resampling process many times, say $10,000$.

3. Construct the bootstrap distribution of the statistics. Inspect its spread, bias, and shape.
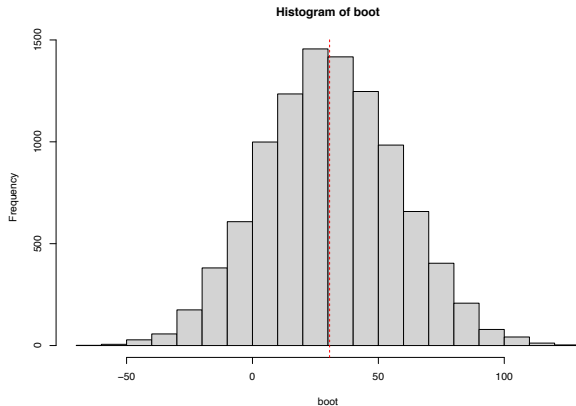
Comparing Two Samples

```r
set.seed(123)
trt <- c(94, 197, 16, 38, 99, 141, 23)
ctrl <- c(52, 104, 146, 10, 51, 30, 40, 27, 46)

boot <- rep(0, 10000)
for(i in 1:10000)
{
  boot.x <- sample(trt, replace = T)
  boot.y <- sample(ctrl, replace = T)
  boot[i] <- mean(boot.x) - mean(boot.y)
}
```

Comparing Two Samples

```
hist(boot) #shape
abline(v = mean(trt) - mean(ctrl), col = "red", lty = 2)
```


Histogram of boot

Comparing Two Samples

```r
# observed value
mean(trt) - mean(ctrl)
```

```
## [1] 30.63492
```

```r
sd(boot) # dispersion
```

```
## [1] 26.89161
```

```r
# Bootstrap estimated Bias
(bias <- (mean(boot) - (mean(trt) - mean(ctrl))))
```

```
## [1] 0.09834603
```

```r
# 95% percentile CI
quantile(boot, c(.025, .975))
```

```
##     2.5%    97.5%
## -20.88889  83.61944
```

```r
sum(boot < 0) / length(boot)
```

```
## [1] 0.1255
```

```r
bias/sd(boot) # accuracy of CI
```

```
## [1] 0.003657127
```

It cannot be ruled out that the treatment could have a negative effect on survival time (12% of the bootstrap values are negative). Thus we would conclude that if a beneficial effect is provided by the treatment, then this benefit is insufficiently strong to be significant in the small sample of data available.

Also, bias is about 0.4% of the standard error.

If the ratio of bias/SE exceeds ± 0.02, then it is large enough to potentially have a substantial effect on the accuracy of confidence intervals.

Your turn

A high school student was curious about the total number of minutes devoted to commercials during any given half-hour time period on basic and extended cable TV channels (B. Rodgers and T. Robinson, personal communication).

ID	Times	Cable
1	7.0	Basic
2	10.0	Basic
3	10.6	Basic
4	10.2	Basic
5	8.6	Basic
6	7.6	Basic
7	8.2	Basic
8	10.4	Basic
9	11.0	Basic
10	8.5	Basic

ID	Times	Cable
11	3.4	Extended
12	7.8	Extended
13	9.4	Extended
14	4.7	Extended
15	5.4	Extended
16	7.6	Extended
17	5.0	Extended
18	8.0	Extended
19	7.8	Extended
20	9.6	Extended

Your turn

```
x.b <- c(7.0, 10.0, 10.6, 10.2, 8.6, 7.6,  8.2, 10.4, 11.0, 8.5)
x.e <- c(3.4, 7.8, 9.4, 4.7, 5.4, 7.6, 5.0, 8.0, 7.8, 9.6)
```

Question: Is there a significant difference in the length of commercials (in minutes) during random half hour periods? Use the **sample()** function.

- Bootstrap the difference in mean times, plot the distribution, and give summary statistics of the bootstrap distribution. Obtain a 95% bootstrap percentile confidence interval, and interpret this interval.

- What is the bootstrap estimate of the bias? What faction of the bootstrap standard error does this represent?

- Use the **boot()** function.

Example

Let's find an approximate 95% confidence interval for theta using the percentile bootstrap approach for the mouse data.

```
boot.ci(b, conf = .95, type = "perc")
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = b, conf = 0.95, type = "perc")
##
## Intervals :
## Level      Percentile
## 95%    (-20.75,  83.98 )
## Calculations and Intervals on Original Scale
```

```
quantile(b$t, c(0.025, 0.975)) # Minor difference due to different algorithms
```

```
##       2.5%      97.5%
## -20.71508   83.98413
```

Example: Mouse data

Histogram of the 10000 Bootstrap Replications
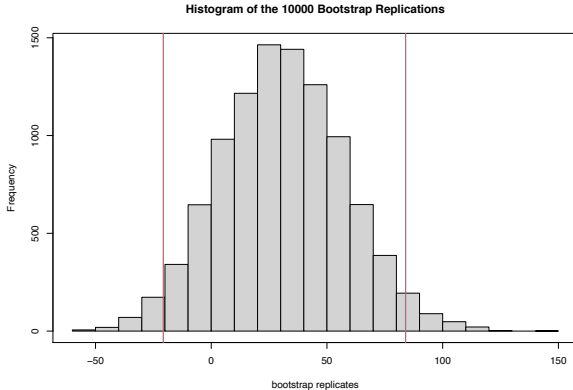
You can also use "infer" or "rasmple" package for this case.

Percentile Bootstrap CI

- The percentile bootstrap CI covers θ with probability less than $1 - \alpha$ when n is small in many cases.
- Coverage improves for large n, but it is possible to develop better intervals by using adjusted percentiles. (BCa Intervals)