Bootstrap

Max Turgeon

STAT 3150–Statistical Computing

Lecture Objectives

- Use bootstrap to estimate the bias and variance of an estimator.
- Understand how the empirical CDF is related to resampling techniques.
- Learn how to compute the different bootstrap confidence intervals.

Motivation

- As with jackknife, the main motivation is to study the sampling distribution of an estimator.
- Jackknife can be used to estimate bias and standard error.
    - But it doesn't always work (e.g. sample median)
- **Bootstrap** is another resampling method that takes a more direct approach to estimating the sampling distribution.

Bootstrap estimate of the standard error i

- Let $X_1, \ldots, X_n$ be a random sample from a distribution F.
- Suppose we use this sample to compute an estimate $\hat{\theta}$ of a population parameter θ.
- Imagine a situation where we can generate B additional samples of size n from the same distribution F.
- For each sample, we could compute an estimate $\hat{\theta}^{(b)}$, where $b = 1, \ldots, B$.
- We could then estimate the *standard error* of $\hat{\theta}$ by taking the *sample standard deviation* of the additional estimates $\hat{\theta}^{(b)}$.
- Of course, we can't really generate these additional samples...

Bootstrap estimate of the standard error ii

- **Bootstrap** mimics this situation by sampling **with replacement** from the original sample $X_1, \ldots, X_n$.
    - Generate a sample $X_1^{(b)}, \ldots, X_n^{(b)}$ of size n by sampling with replacement from the original sample.
    - Compute $\hat{\theta}^{(b)}$ using that bootstrap sample.

Example i

- Let's revisit the sample median but with bootstrap
  - Recall that the jackknife estimate of the standard error was too small

```r
population <- seq(1, 100)
median(population)
```

```
## [1] 50.5
```

Example ii

```r
# Generate B samples from sampling distribution
B <- 5000
n <- 10
results <- replicate(B, {
    some_sample <- sample(population,
                          size = n)
    median(some_sample)
})
sd(results)
```

```
## [1] 13.04957
```

Example iii

```r
# Take a single sample from population
one_sample <- sample(population, size = n)
median(one_sample)
```

```
## [1] 28.5
```

```r
# How do we sample with replacement?
sample(n, n, replace = TRUE)
```

```
##  [1]  4 10  1  8  4  4  4  7  8  2
```

Example iv

```r
# Bootstrap estimate of SE
boot_theta <- replicate(5000, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  median(one_sample[indices])
})
sd(boot_theta)
```

```
## [1] 8.14544
```

Example v

```r
# Compare with jackknife
theta_hat <- median(one_sample)
theta_i <- numeric(n)
for (i in 1:n) {
    theta_i[i] <- median(one_sample[-i])
}
sqrt((n-1)*mean((theta_i - mean(theta_i))^2))
```

```
## [1] 1.5
```

Example i

- We will revisit the `law` dataset in the `bootstrap` package, which contains information on average `LSAT` and `GPA` scores for 15 law schools.
- We are interested in the correlation ρ between these two variables

```
library(bootstrap)
# Estimate of rho
(rho_hat <- cor(law$LSAT, law$GPA))
```

```
## [1] 0.7763745
```

Example ii

```r
# Bootstrap estimate of SE
n <- nrow(law)
boot_rho <- replicate(5000, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  # We're sampling pairs of observations
  # to keep correlation structure
  cor(law$LSAT[indices], law$GPA[indices])
})
```

```r
sd(boot_rho)
```

Example iii

```
## [1] 0.1360481
```

Empirical CDF i

- We briefly mentioned the empirical CDF in the module on Data Visualization.
- More formally, the **empirical CDF** of a sample $X_1, \ldots, X_n$, denoted $\hat{F}_n$, is the CDF of a *discrete* distribution whose support is the data points $\{X_1, \ldots, X_n\}$, and where each point has mass $1/n$.
- Mathematically, we have

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq x).$$

Empirical CDF ii

- **Why do we care?** We already argued that we can't easily generate more samples from F. Instead, bootstrap generates more samples from the distribution $\hat{F}_n$.
    - Sampling with replacement is the same as sampling from the empirical CDF!
    - Since $\hat{F}_n \to F$, we can often translate this convergence in terms of the bootstrap estimates.

$$\begin{array}{rccccc} \text{Real world:} & F & \Rightarrow & X_1, \ldots, X_n & \Rightarrow & \hat{\theta} = g(X_1, \ldots, X_n) \\ \text{Bootstrap world:} & \hat{F}_n & \Rightarrow & X_1^{(b)}, \ldots, X_n^{(b)} & \Rightarrow & \hat{\theta}^{(b)} = g(X_1^{(b)}, \ldots, X_n^{(b)}) \end{array}$$

Bootstrap estimateof bias

- Just as with jackknife, we can use bootstrap to estimate the bias of $\hat{\theta}$.
- Let $\hat{\theta}^{(b)}$ be the estimates computed using the bootstrap samples, and let $\bar{\theta} = n^{-1} \sum_{b=1}^{B} \hat{\theta}^{(b)}$ be their sample mean.
- The **bootstrap estimate of bias** is given by

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

Example i

```r
# law dataset
rho_hat <- cor(law$LSAT, law$GPA)

# Bootstrap estimate of bias
B <- 5000
n <- nrow(law)
```

Example ii

```r
boot_rho <- replicate(5000, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  # We're sampling pairs of observations
  # to keep correlation structure
  cor(law$LSAT[indices], law$GPA[indices])
})

(bias <- mean(boot_rho) - rho_hat)
```

```
## [1] -0.004382551
```

Example iii

```r
# Debiased estimate
rho_hat - bias
```

```
## [1] 0.780757
```

Bootstrap confidence intervals

- There are several ways to construct confidence intervals in bootstrap:
    - Standard normal bootstrap
    - Bootstrap percentile
    - Basic bootstrap
    - Student bootstrap
- They all have different properties, and they can all be useful depending on the context.

Standard normal bootstrap CI i

- This is similar to what we've been doing until now.
- It relies on the Central Limit Theorem:

$$\frac{\hat{\theta} - E(\hat{\theta})}{SE(\hat{\theta})} \to N(0, 1).$$

- If we estimate $\widehat{bias}(\hat{\theta})$ and $SE(\hat{\theta})$ using bootstrap, then we can construct an approximate $100(1-\alpha)$% confidence interval for θ via

$$\hat{\theta} - \widehat{bias}(\hat{\theta}) \pm z_{\alpha/2} SE(\hat{\theta}).$$

Standard normal bootstrap CI ii

- This interval is easy to compute, but it assumes that the sampling distribution is approximately normal.
    - Works well for estimators $\hat{\theta}$ that can be expressed as a sample mean (e.g. Monte Carlo integration)
    - Doesn't work well when the sampling distribution is skewed.

Bootstrap percentile CI

- Let $\hat{\theta}^{(b)}$, $b = 1, \ldots, B$ be the bootstrap estimates.
- The **bootstrap percentile confidence interval** is the interval of the form $(\hat{\theta}_{\alpha/2}, \hat{\theta}_{1-\alpha/2})$, where $\hat{\theta}_{\alpha/2}$ and $\hat{\theta}_{1-\alpha/2}$ are the $\alpha/2$-th and $1-\alpha/2$-th sample quantiles of the bootstrap estimates, respectively.
- This is also very simple to compute, and it will account for the skewness in the sampling distribution.

Basic bootstrap CI i

- This is also known as the **pivotal bootstrap CI**.
- It is very similar to the bootstrap percentile approach, but instead of taking the sample quantiles of $\hat{\theta}^{(b)}$, $b = 1, \ldots, B$, we take the sample quantiles of the *pivot quantities* $\hat{\theta}^{(b)} - \hat{\theta}$, $b = 1, \ldots, B$.
- Note that the β-th quantile of $\hat{\theta}^{(b)} - \hat{\theta}$ is equal to $\hat{\theta}_\beta - \hat{\theta}$, where $\hat{\theta}_\beta$ is the β-th quantile of $\hat{\theta}^{(b)}$.
- To build the basic bootstrap CI, we take $\hat{\theta}$ minus some critical values. But instead of using the critical values of the standard normal, we take our critical values from the *pivot quantities*:

$$\hat{\theta} - (\hat{\theta}_\beta - \hat{\theta}) = 2\hat{\theta} - \hat{\theta}_\beta.$$

Basic bootstrap CI ii

- Therefore, the **basic bootstrap** $100(1-\alpha)\%$ confidence interval for θ is

$$(2\hat{\theta} - \hat{\theta}_{1-\alpha/2}, 2\hat{\theta} - \hat{\theta}_{\alpha/2}).$$

- **Why use basic over percentile?** It turns out the basic bootstrap CI has better theoretical properties and stronger convergence guarantees.

Student bootstrap CI i

- This confidence interval accounts for the fact we have to estimate the standard error.
- However, it is much more involved: we can construct an approximate $100(1-\alpha)\%$ confidence interval for θ via

$$\left(\hat{\theta} - t^*_{1-\alpha/2} SE(\hat{\theta}), \hat{\theta} - t^*_{\alpha/2} SE(\hat{\theta})\right),$$

  where $t^*_{1-\alpha/2}$ and $t^*_{\alpha/2}$ are computed using a **double bootstrap**, and where $SE(\hat{\theta})$ is the usual bootstrap estimate of the standard error.

Student bootstrap CI ii

Algorithm

1. For each bootstrap sample estimate $\hat{\theta}^{(b)}$, compute a "t-type" statistic $t^{(b)} = \frac{\hat{\theta}^{(b)} - \hat{\theta}}{SE(\hat{\theta}^{(b)})}$, where $SE(\hat{\theta}^{(b)})$ is specific to the b-th sample, and it can be computed using bootstrap on the samples $X_1^{(b)}, \ldots, X_n^{(b)}$.
2. From the sample $t^{(b)}, b = 1, \ldots, B$, let $t^*_{1-\alpha/2}$ and $t^*_{\alpha/2}$ be the $1 - \alpha/2$-th and $\alpha/2$-th sample quantiles.

This confidence interval is more accurate than the standard normal bootstrap CI, but this accuracy comes with a large computational cost.

Example i

We will compute all four types of confidence intervals for the correlation between LSAT and GPA scores.

Example ii

```r
B <- 5000
n <- nrow(law)
boot_rho <- replicate(B, {
  # Sample with replacement
  indices <- sample(n, n, replace = TRUE)
  cor(law$LSAT[indices], law$GPA[indices])
})

rho_hat <- cor(law$LSAT, law$GPA)
bias <- mean(boot_rho) - rho_hat
se <- sd(boot_rho)
```

Example iii

```r
# 1. Standard normal
c(rho_hat - bias - 1.96*se,
  rho_hat - bias + 1.96*se)
```

```
## [1] 0.5226827 1.0446077
```

```r
# 2. Bootstrap percentile
quantile(boot_rho,
         probs = c(0.025, 0.975))
```

```
##      2.5%     97.5%
## 0.4610949 0.9637334
```

Example iv

```r
# 3. Basic bootstrap
crit_vals <- quantile(boot_rho,
                      probs = c(0.025, 0.975))
c(2*rho_hat - crit_vals[2],
  2*rho_hat - crit_vals[1],
  use.names = FALSE)
```

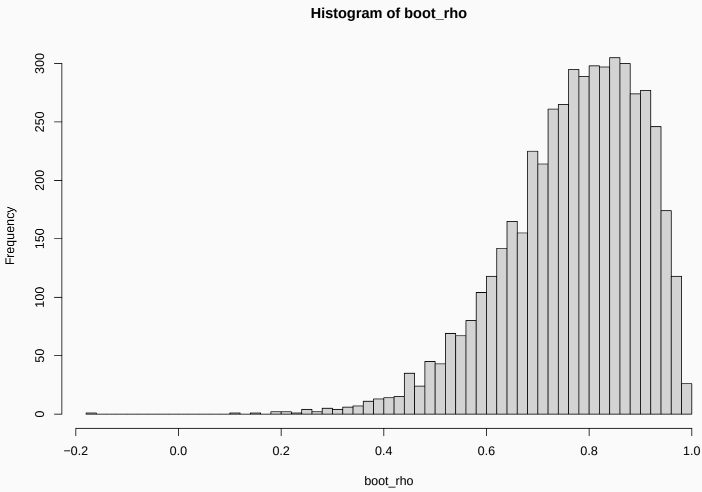
```
## [1] 0.5890155 1.0916540
```

Example v

Table 1: Only the percentile method gives a sensible confidence interval, i.e. a CI that is contained within the interval $(-1, 1)$.

Method	95% CI
Standard Normal	(0.52, 1.04)
Percentile	(0.46, 0.96)
Basic Bootstrap	(0.59, 1.09)

Example vi


Histogram of boot_rho
Frequency
0
50
100
150
200
250
300
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
boot_rho

Example vii

```r
# 4. Student bootstrap
boot_rho_t <- replicate(B, {
  indices <- sample(n, n, replace = TRUE)
  rho_b <- cor(law$LSAT[indices], law$GPA[indices])
  double_boot <- replicate(100, {
    double_ind <- sample(indices, n, replace = TRUE)
    cor(law$LSAT[double_ind], law$GPA[double_ind])
  })
  tb <- (rho_b - rho_hat)/sd(double_boot)
  return(c(rho_b, tb))
})
```

Example viii

```r
# The output has two rows:
# First row: rho_b values
# Second row: tb values
str(boot_rho_t)
```

```
## num [1:2, 1:5000] 0.837 0.551 0.755 -0.16
0.823 ...
```

```r
# SE estimated using rho_b values
SE <- sd(boot_rho_t[1,])
```

Example ix

```r
# t critical values
tcrit_vals <- quantile(boot_rho_t[2,],
                       probs = c(0.025, 0.975))
```

```r
c(rho_hat - tcrit_vals[2]*SE,
  rho_hat - tcrit_vals[1]*SE,
  use.names = FALSE)
```

```
## [1] -0.2508246  0.9874933
```

Example x

- This is a valid confidence interval, but it is much wider than the other three!
- Given the skewness of the bootstrap samples, the percentile approach is the most appropriate.

Final remarks

- So when should we use jackknife vs bootstrap?
- In some way, the jackknife is an *approximation* of the bootstrap, and as a consequence, the bootstrap almost always outperforms the jackknife.
- However, for small sample sizes, the jackknife will be more computationally efficient:
    - Jackknife requires $n + 1$ computations of the estimate.
    - Bootstrap requires $B + 1$ computations of the estimate, where B is usually at least 1000.
- Bootstrap performs better when the sampling distribution is skewed.
- Jackknife does **not** work with some estimators, e.g. sample median and sample quantiles.