Resampling
Bootstrap Method

1. Introduction
2. The Bootstrap Method
3. Bootstrap Estimation of Standard Error
4. Bootstrap Estimation of Bias
5. Bootstrap Confidence Interval

1 Introduction

2 The Bootstrap Method

3 Bootstrap Estimation of Standard Error

4 Bootstrap Estimation of Bias

5 Bootstrap Confidence Interval

Introduction

- In the chapter of simulation, we learn how we can do simulations when the underlying distributions are known.
- What if the underlying distributions are not known (all we have is just the data)?

Example (Law School)

A random sample of 15 law schools was selected. The average score on law school admission test (LSAT) and the average undergraduate grade-point average (GPA) for each school were recorded in a file, `lawschool.csv`.
We are interested in the correlation coefficient ρ, which can be estimated by the sample correlation coefficient r. Find the estimate of the standard error of r.

- Note that, we have small data and the underlying distribution of LSAT and SAT both are unknown.
- We could use (Nonparametric) Bootstrap Method to estimate the standard error of r!

Bootstrap Method (Intro)

- The (nonparametric) Bootstrap Method was introduced in 1979 by Efron.
- It is a class of nonparametric Monte Carlo methods that estimate the distribution of an estimator by resampling.
- Resampling methods treat an observed sample as a finite population.
- Random samples are generated or resampled from the observed/original sample.
- These random samples are used to estimate population characteristics and make inferences about the sampled population.
- Non-parametric bootstrap methods are often used when the distribution of the target population is not specified (hence the name nonparametric); the sample is the only information available.
- The distribution of the finite population represented by the sample can be regarded as a pseudo-population with similar characteristics as the true population.

Difference Between Simulation and Bootstrap

- Simulation generates samples from completely specified distribution.

- Parametric bootstrap: fits/estimates a distribution for the given sample, $f(x, \alpha)$, and then generates random samples from this fitted distribution.

- Nonparametric bootstrap: does not fit any distribution to the given sample, just generates random samples from the empirical distribution of the sample.
  - Empirical distribution:

$$f_n(x) = \begin{cases} 1/n, & x = x_1, x_2, x_3, ..., x_n; \\ 0, & \text{otherwise} \end{cases}$$

  - Empirical cumulative distribution

$$F_n(t) = P(x \leq t) = \frac{\text{number of } x\text{'s} \leq t}{n}.$$

1 Introduction

2 The Bootstrap Method

3 Bootstrap Estimation of Standard Error

4 Bootstrap Estimation of Bias

5 Bootstrap Confidence Interval

Typically

- The BM is typically is used to find
  - Standard errors for estimators;
  - confidence intervals for unknown parameters;
  - p-values for test statistics under a null hypothesis.
- It helps to estimate quantities associated with the sampling distribution of estimators and test statistics.
- Useful when standard assumptions invalid, e.g. n small, data not normal.

The Bootstrap Method (BM)

- Suppose θ is the parameter of interest (θ could be a vector), and $\hat{\theta}$ is an estimator of θ.

  For example:
  - θ could be the population mean μ and $\hat{\theta}$ could be $\bar{X}$.
  - θ could be the population correlation between two variables, ρ, and $\hat{\theta}$ could be the sample correlation from a random sample, r.

- We would want to estimate the sampling distribution of the estimators, $F_{\hat{\theta}}$. BM is used in the estimation steps to derive the bootstrap estimate of $F_{\hat{\theta}}$.

Steps of the Bootstrap Estimation

- (A) For each bootstrap replicate, indexed $b = 1, 2, ..., B$:

  A.1 generate bootstrap sample $x^{*(b)} = x_1^*, x_2^*, ..., x_n^*$ by sampling with replacement from the observed sample $x_1, x_2, ..., x_n$. This is the nonparametric part. This step is different for parametric boostrap in slide 11.

  A.2 compute the value of the estimator from bth bootstrap sample $x^{*(b)}$, which is denoted as $\hat{\theta}^{*(b)}$.

- (B) At the end of (A) , we have

$$\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, ..., \hat{\theta}^{*(B)}.$$

  The boostrap estimate (BE) of $F_{\hat{\theta}}$ is then the empirical distribution of these replicates.

- (C) The BE $F_{\hat{\theta}}$ is used to estimate the standard error, bias and confidence interval of an estimator (in the following sections).

Notes on Parametric Bootstrap

- When the distrbution of the population (where sample was collected) is unknown, we might estimate that distribution from the observed sample, say $f_X(x, \alpha)$.

- In the step A.1 of nonparametric bootstrap, we replace sampling with replacement from original sample by sampling from $f_X(x, \alpha)$.

- For example, we estimate that the sample was collected from a population with distribution $f_X(x, \alpha)$ where α is the parameter (could be a vector).
  - We then estimate α by $\hat{\alpha}$ based on the observed sample $x_1, x_2, ..., x_n$. One could use MLE at this step.

- We then generate the bootstrap sample $x^{*(b)} = x_1^*, x_2^*, ..., x_n^*$, $b = 1, ..., B$ by simulating from $f_X(x, \hat{\alpha})$.

1 Introduction

2 The Bootstrap Method

3 Bootstrap Estimation of Standard Error

4 Bootstrap Estimation of Bias

5 Bootstrap Confidence Interval

Standard Error of an Estimator in General

- Variables $X_1, X_2, ..., X_n$ has the observed values as $x_1, x_2, ..., x_n$.

- θ is the parameter of interest. Its estimator is $\hat{\theta}(X_1, X_2, ..., X_n)$ which is a function of $X_1, X_2, ..., X_n$.

- From the observed sample, an estimate value of θ is $\hat{\theta}(x_1, x_2, ..., x_n)$. We would want to estimate the SE of this estimation.

Bootstrap Estimation of SE of an Estimator

- The bootstrap estimate of the SE **is the sample standard deviation** of the bootstrap replicates $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, ..., \hat{\theta}^{*(B)}$, which is

$$\hat{\mathsf{se}}(\hat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\theta}^{*(b)} - \overline{\hat{\theta}^{*}})^2}$$

where $\overline{\hat{\theta}^{*}} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$.

Law School Example (1)

Example of Law School in slide 4.

```
> law = read.csv("C:/Data/lawschool.csv"); law
   LSAT  GPA
1   576 3.39
2   635 3.30
3   558 2.81
4   578 3.03
5   666 3.44
6   580 3.07
7   555 3.00
8   661 3.43
9   651 3.36
10  605 3.13
11  653 3.12
12  575 2.74
13  545 2.76
14  572 2.88
15  594 2.96
```

Law School Example (2)

```
> attach(law)
> cor(LSAT,GPA) # r = 0.776
[1] 0.7763745
> #set.seed(999)
> # NONPARAMETRIC BOOTSTRAP
> R <- 1000 # number of bootstrap replicates;
> n <- length(GPA) # sample size
> theta.b <- numeric(R) # storage for boostrap estimates
> for (b in 1:R) {
+ # for each b, randomly select the indices, sampling with replaceme
+     i <- sample(1:n, size=n, replace=TRUE)
+     LSATb <- LSAT[i] # i is a vector of indices
+     GPAb <- GPA[i]
+     theta.b[b] <- cor(LSATb, GPAb)
+ }
> sd(theta.b)
[1] 0.1395574
```

So the bootstrap estimate of the standard error of r is as the output above.

The boot Function in R

```
> library(boot)
> bcor <- function(data, bindex){
+   return(cor(data[bindex,1], data[bindex,2]))
+ }
> boot.cor <- boot(law, statistic=bcor, R=1000)
> # Obtain the bias and standard error
> boot.cor
```

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = law, statistic = bcor, R = 1000)


Bootstrap Statistics :
     original       bias    std. error
t1* 0.7763745 -0.005277963   0.1366259
```

```
https://cran.r-project.org/web/packages/boot/boot.pdf
```

1 Introduction

2 The Bootstrap Method

3 Bootstrap Estimation of Standard Error

4 Bootstrap Estimation of Bias

5 Bootstrap Confidence Interval

Bootstrap Estimation of Bias

- $\hat{\theta}$ is the estimator of θ.

- The bias of the estimator $\hat{\theta}$ for θ is:

$$\text{bias}(\hat{\theta}) = E[\hat{\theta} - \theta] = E[\hat{\theta}] - \theta.$$

- The boostrap estimate of the bias of an estimator $\hat{\theta}$ **is the difference** between the mean of the boostrap replicates $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, ..., \hat{\theta}^{*(B)}$ and $\hat{\theta}$, i.e.,

$$\widehat{\text{bias}}(\hat{\theta}) = \overline{\hat{\theta}}^{*} - \hat{\theta}$$

where $\overline{\hat{\theta}^{*}} = \dfrac{1}{B}\sum_{b=1}^{B}\hat{\theta}^{*(b)}$ and $\hat{\theta}$ is the estimate computed from the original sample.

Law School Example

Boostrap estimate of bias

```
> theta.hat = cor(LSAT,GPA) # value computed from original sample
> B <- 1000 # n = 15 from previous code
> theta.b <- numeric(B)# storage for boostrap estimates
> for (b in 1:B) {
+   i <- sample(1:n, size=n, replace=TRUE)
+   LSATb <- LSAT[i]
+   GPAb <- GPA[i]
+   theta.b[b] <- cor(LSATb, GPAb)
+   }
> bias <- mean(theta.b)- theta.hat
> bias
[1] -0.005170202
```

- Alternatively, we can have the result from `boot` function.

1 Introduction

2 The Bootstrap Method

3 Bootstrap Estimation of Standard Error

4 Bootstrap Estimation of Bias

5 Bootstrap Confidence Interval

Some Types of Bootstrap Confidence Interval

- The basic bootstrap CI
- The percentile bootstrap CI
- The normal bootstrap CI
- The studentized bootstrap CI
- The adjusted bootstrap percentile CI

We introduce the first three types.

The Basic Bootstrap Confidence Interval

- The quantiles of the bootstrap samples are used to determine the confidence limits.

- The $100(1-\alpha)\%$ confidence limits for the basic bootstrap confidence interval are

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

  where $\hat{\theta}^*_{\alpha/2}$ is the α sample quantile from the empirical distribution function of the replicates $\hat{\theta}^*$.

- A 95% basic bootstrap CI for the correlation coefficient in the Law School is presented as an example.

The Basic Bootstrap CI for Law School Example

```
> R = 2000 # larger for estimating confidence interval
> theta.b = numeric(R)
> alpha = 0.05; CL = 100*(1-alpha)
> for (b in 1:R) {
+ i <- sample(1:n, size=n, replace=TRUE)
+ LSATb <- LSAT[i]
+ GPAb <- GPA[i]
+ theta.b[b] <- cor(LSATb, GPAb)
+ }
```

```
> low = quantile(theta.b, alpha/2)
> high = quantile(theta.b, 1 - alpha/2)
> cat("A",CL,"% basic confidence interval is ",
+     2*theta.hat - high, 2*theta.hat - low,"\n")
A 95 % basic confidence interval is  0.5936548 1.107248
```

The Percentile Bootstrap Confidence Interval

- $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, ..., \hat{\theta}^{*(B)}$ are bootstrap replicates of the statistics $\hat{\theta}$.

- From the empirical distribution function of the replicates, compute the $\alpha/2$ quantile $\hat{\theta}^*_{\alpha/2}$ and the and the $1-\alpha/2$ quantile $\hat{\theta}^*_{1-\alpha/2}$.

- The $100(1-\alpha)\%$ percentile bootstrap CI for θ is defined as

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

The Percentile Bootstrap CI for Law School Example

```
> low <- quantile(theta.b, alpha/2)
> high <- quantile(theta.b, 1-alpha/2)
> CL <- 100*(1-alpha)
> cat("A",CL,"% bootstrap CI is", low, high,"\n")
A 95 % bootstrap CI is 0.4455005 0.9590942
```

The Normal Bootstrap Confidence Interval

- The normal bootstrap CI constructs the CI based on the assumption that the distribution of the estimator is normally distributed.

$$\hat{\theta} \sim N(\theta + \text{bias}, \text{variance})$$

where we then can estimate θ by the value of $\hat{\theta}$ form the original sample. `bias` is estimated using bootstrap replicates $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, ..., \hat{\theta}^{*(B)}$, and `variance` is the sample variance of $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, ..., \hat{\theta}^{*(B)}$.

- The $100(1-\alpha)\%$ normal bootstrap CI for θ is then defined as

$$\left(\hat{\theta} - \text{bias} \pm z_{(1-\alpha/2)} \times \sqrt{\text{variance}}\right).$$

The Normal Bootstrap CI for Law School Example

```
> bias = mean(theta.b) - theta.hat
> se = sd(theta.b)
> low <- theta.hat - bias - 1.96*se
> high <- theta.hat - bias + 1.96*se
> cat("A",CL,"% bootstrap CI is",
+     low, high,"\n")
A 95 % bootstrap CI is 0.5183496 1.05845
```

Bootstrap Confidence Interval by boot.ci

```
> library(boot)
> bcor <- function(data, bindex){
+ return(cor(data[bindex,1], data[bindex,2]))
+ }
> boot.cor <- boot(law, statistic=bcor, R=2000)
```

To get all three types of CI, we specify the 3 types.

```
> boot.ci(boot.cor,type=c("basic","perc","norm"))
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 2000 bootstrap replicates

CALL :
boot.ci(boot.out = boot.cor, type = c("basic", "perc", "norm"))

Intervals :
Level      Normal              Basic               Percentile
95%   ( 0.5291,  1.0399 )   ( 0.5941,  1.0718 )   ( 0.4809,  0.9587
Calculations and Intervals on Original Scale
```