Chapter 5 - Beyond Monte Carlo

Bootstrap Method for Bias and Standard Deviation Estimation.

Prof. Alex Alvarez, Ali Raisolsadat

School of Mathematical and Computational Sciences
University of Prince Edward Island

Framework

Sample Data: $X = (X_1, X_2, \ldots, X_n)$ independent, identically distributed with $X_i \sim F$, where F is unknown. Usually we observe a realization $x = (x_1, x_2, \ldots, x_n)$ of X.

We are interested in some statistic $\theta = \theta(F)$.

Assume that we have some estimator $\widehat{\theta}_n = \widehat{\theta}_n(X)$ of θ.

Today we will focus on the study of two problems:

- Estimation of the bias of $\widehat{\theta}_n$
- Estimation of the standard deviation of $\widehat{\theta}_n$

Formulation

In our previous lecture we defined the empirical cumulative distribution function as:

$$F_n(t) = \frac{1}{n}\sum_{i=1}^{n} 1_{(-\infty,t]}(x_i)$$

and we saw that the approximation $F_n \approx F$ may be valid even for not so large values of n.

We will generate Bootstrap samples $x^{*(j)} = \left(x_1^{*(j)}, x_2^{*(j)}, ...x_n^{*(j)}\right)$ with $j = 1, 2, ...N$ where the terms $x_m^{*(j)}$ are i.i.d. random values distributed according to F_n.

An estimator $\widehat{\theta}_n = \widehat{\theta}_n(X)$ of θ satisfies the **plug-in principle** if the relation

$$\widehat{\theta}_n(x) = \theta(F_n)$$

is valid for all $x = (x_1, x_2, ..., x_n)$.

Bias Estimation

We know that $bias\left(\widehat{\theta}_n\right) = E(\widehat{\theta}_n) - \theta = E_F(\widehat{\theta}_n) - \theta(F)$

As we do not know F, but we do know F_n, we will approximate $bias\left(\widehat{\theta}_n\right)$ with

$$E_{F_n}(\widehat{\theta}_n) - \theta(F_n)$$

If θ_n satisfies the plug-in principle, we can replace $\theta(F_n)$ with $\widehat{\theta}_n(x)$. For the estimations of $E_{F_n}(\widehat{\theta}_n)$ we will use the sample mean of the Bootstrap sample estimates $\widehat{\theta}_n(x^{*(j)})$ with $j = 1, 2, ...N$. Then we will have:

$$\widehat{bias}\left(\widehat{\theta}_n\right) = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_n\left(x^{*(j)}\right) - \widehat{\theta}_n(x)$$

Bootstrap Bias Estimation Example

The following data are students grades in some course. The grades are distributed according to some unknown distribution. Use the sample median to estimate the median of the unknown distribution and estimate the bias of your estimator.

Grades: $x = (84, 63, 92, 89, 80, 47, 68, 66, 56)$

Solution: By ordering the data in ascending order (47, 56, 63, 66, 68, 80, 84, 89, 92) we can see that the sample median is $\widehat{\theta}_9(x) = 68$.

For the other term in the bias estimation, we will generate $N = 1000$ Bootstrap samples as in the following code.

Bootstrap Bias Estimation Example

R Code

```r
x <- c(84, 63, 92, 89, 80, 47, 68, 66, 56)
N <- 1000
median_vector <- vector()
for(i in c(1:N)) {
 Y <- sample(x, size=9, replace=TRUE)
 median_vector[i] = median(Y)
}

average_medians <- mean(median_vector)
```

By running this code once we get

$$\widehat{bias}\left(\widehat{\theta}_n\right) = \frac{1}{N}\sum_{j=1}^{N}\widehat{\theta}_n\left(x^{*(j)}\right) - \widehat{\theta}_n(x) = 71.34 - 68 = 3.34$$

Estimation of standard deviation

The problem of estimating the standard deviation of the estimator $\widehat{\theta}_n$ under the (unknown) distribution F is simplified by considering the standard deviation under the (known) empirical distribution F_n:

$$s.d._F\left(\widehat{\theta}_n\right) \approx s.d._{F_n}\left(\widehat{\theta}_n\right)$$

Then, we approximate $s.d._{F_n}\left(\widehat{\theta}_n\right)$ by the empirical standard deviation of the Bootstrap estimates $\widehat{\theta}_n\left(x^{*(j)}\right)$ with $j = 1, 2, ...N$

The code corresponding to the estimation of the standard deviation of the median estimator in the previous example is given as follows.

Bootstrap Bias Estimation Example

R Code

```r
x <- c(84, 63, 92, 89, 80, 47, 68, 66, 56)
N <- 1000
median_vector <- vector()
for (i in c(1:N)) {
  Y <- sample(x, size=9, replace=TRUE)
  median_vector[i] = median(Y)
}
sd_vector <- sd(median_vector)
```

By running this code once we get that $\widehat{s.d}(\widehat{\theta}_n) = 8.38$.

Homework:

With the data from the previous example, find the Bootstrap estimate of the standard deviation of the **sample mean** estimator.