

Bootstrap: introduction (1/2)

Suppose that we observe a sample x_1, \dots, x_n which are realizations of i.i.d. r.v. X_1, \dots, X_n , with distribution L_x and with CDF F_x unknown.. The empirical Cumulative Distribution Function (eCDF) is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{x_i \leq x}$$

and is an estimator of the true CDF F_x . Moreover, we have that

$$E[F_n(x)] = F_x$$

and

$$\text{var}(F_n(x)) = \frac{1}{n} \left((F_x)(1 - F_x) \right)$$

Bootstrap: introduction (2/2)

Besides, from the Law of Large Numbers (LLN), we have that

$$F_n(x) \xrightarrow{a.s.} F_x \quad \text{as } n \text{ goes to } \infty.$$

And from the Central Limit Theorem (CLT), we have that

$$F_n(x) \xrightarrow{L} N\left(F_x, \frac{1}{n}F_x(1 - F_x)\right) \quad \text{as } n \text{ goes to } \infty.$$

The underlying idea of nonparametric bootstrap is that bootstrap samples can be generated from the $F_n(x)$ when F_x is unknown since $F_n(x)$ is a consistent estimator for F_x .

General algorithm for Bootstrap

1. B independent samples $x_{b1}^*, \dots, x_{bn}^*$, $b = 1, \dots, B$ are drawn with replacement from the data x_1, \dots, x_n .
2. On each of the B independent samples, an estimate $\hat{\theta}_b^*$ is computed (the bootstrap replicates of $\hat{\theta}$).
3. The expectation of $\hat{\theta}$ (given $F_n(x)$) is estimated as follows:

$$E_{boot}[\hat{\theta}] = \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^*$$

4. The variance of $\hat{\theta}$ (given $F_n(x)$) is estimated as follows:

$$var_{boot}(\hat{\theta}) = \frac{1}{B-1} \sum_{b=1}^B \left(\hat{\theta}_b^* - \frac{1}{B} \sum_{b=1}^B \hat{\theta}_b^* \right)^2$$

Example 1

Example: We observe a the following sample:

$$\begin{aligned}x_1, \dots, x_n = & 8.26, 6.33, 10.4, 5.27, 5.35, 5.61, 6.12, \\ & 6.19, 5.2, 7.01, 8.74, 7.78, 7.02, 6, 6.5, \\ & 5.8, 5.12, 7.41, 6.52, 6.21, 12.28, 5.6, 5.38, 6.6, 8.74\end{aligned}$$

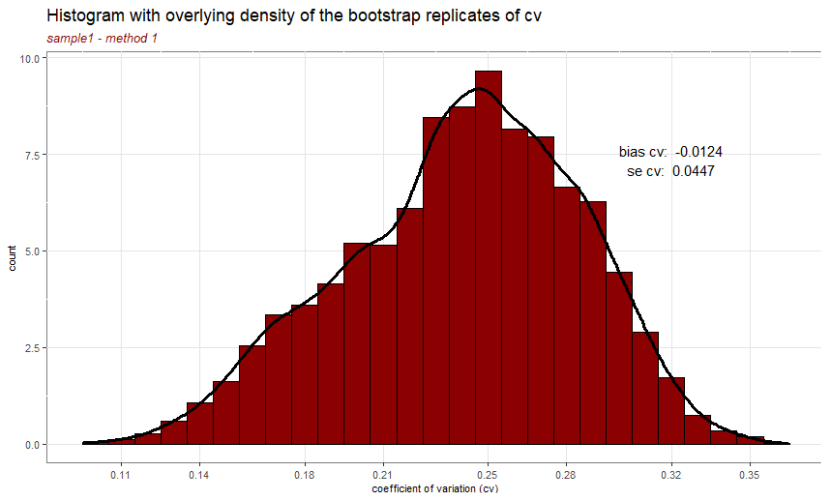
Estimate the bias and the standard error of $\hat{\theta}$, the estimate of the coefficient of variation. The coefficient of variation cv is equal to

$$cv = sd(x)/mean(x)$$

Method 1 the R language

```
1 library(bootstrap)
2
3 sample1 = c(8.26, 6.33, 10.4, 5.27, 5.35, 5.61, 6.12, 6.19, 5.2, 7.01, 8.74,
4 7.78, 7.02, 6, 6.5, 5.8, 5.12, 7.41, 6.52, 6.21, 12.28, 5.6, 5.38, 6.6, 8.74)
5
6 #-----
7 # Perform bootstrap to estimate the cv - mehthod 1
8 #-----
9
10 # define the function to estimatee the cv
11 theta_hat = function(x) {
12   sd(x) / mean(x) # coefficient of variation
13 }
14
15 set.seed(2023)
16 B = 10000
17 boot = bootstrap(x = sample1,          # x: initial sample
18                 n = B,                 # n: number of bootstrap replicates,
19                 theta_hat)             # the function
20
21 # the bootstrap replicates are saved in 'thetastat'
22 # which is an object returned by the function 'bootstrap'
23
24 # estimate of the bias
25 bias <- mean(boot$thetastar) - theta_hat(sample1) # the bias estimate
26 bias # [1] -0.01236049
27
28 # estimate of the standard error
29 sdthetastar <- sd(boot$thetastar)
30 sdthetastar # [1] 0.04466796
```

Method 1: histogram of replicates of cv



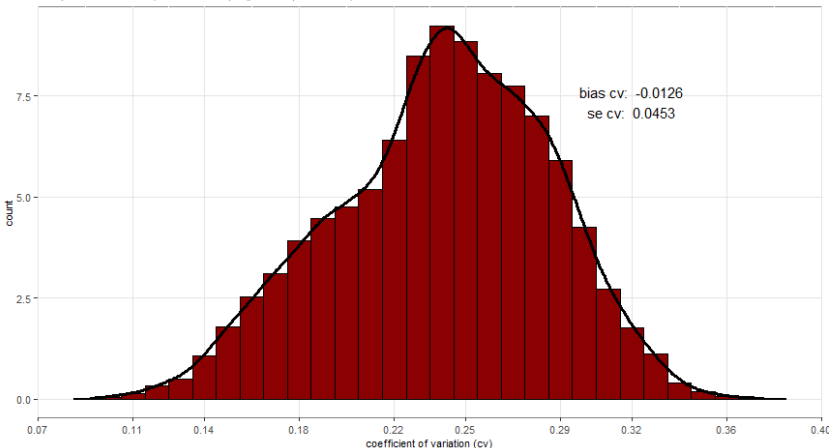
Method 1 the R language

```
1 #-----
2 # Perform bootstrap to estimate the cv - mehthod 2
3 #-----
4 set.seed(2023)
5 B = 10000                                     # B: set the number of
        bootstrap replicates
6 bootstrap_object <- matrix(rep(0, B*length(sample1)),
7                             nrow = B)
8 cv <- numeric(B)
9
10 # perform the bootstrap using the function sample() with replacement
11 for(i in 1:B) {
12
13   bootstrap_object[i,] <- sample(sample1, size = length(sample1), replace = TRUE
14   )
15   cv[i] <- sd(bootstrap_object[i, ]) / mean(bootstrap_object[i, ])
16 }
17
18 # estimate of the bias
19 bias <- mean(cv) - (sd(sample1)/mean(sample1))
20 bias # [1] -0.01255868
21
22 # estimate of the standard error
23 sd <- sd(cv)
24 sd # [1] 0.0452905
```

Method 1: histogram of replicates of cv

Histogram with overlying density of the bootstrap replicates of cv

sample1 - method 2 (function sample() with replacement)



References

Rizzo, M.L. (2019). Statistical Computing with R, Second Edition (2nd ed.). Chapman and Hall/CRC.

<https://doi.org/10.1201/9780429192760>

The R Project for Statistical Computing:

<https://www.r-project.org/>

course notes