Toulouse
School of
Economics

Worksheet 1 – Bootstrap

High-Dimensional Data Analysis and Machine Learning

Camille Mondon

February 3, 2025

This worksheet[1] illustrates the use of the bootstrap in various estimation problems: building confidence intervals for a mean (Exercise 1), estimating the variance of the sample mean (Exercise 2), and building confidence intervals for parameters in linear regression (Exercise 3).

The simulations use `R`. Recall that you can use the `?` command to learn how to use a given `R` function. For instance, you can type

```
?rgamma
```

to show the help page for the function `rgamma`.

1 A first bootstrap estimation

Exercise 1. In this exercise, we use the bootstrap to obtain a confidence interval for the mean.

1. Using the function `rgamma`, generate a random sample of size $n = 15$ from the $\Gamma(2,1)$ distribution[2]. What is the resulting sample mean? Is it close to the corresponding population mean? (namely, the mean of the $\Gamma(2,1)$ distribution)?

2. Create a function that

   - takes a given sample (vector) as input,
   - generates a bootstrap sample from this sample (use the function `sample` to do this), and
   - returns the sample mean of this bootstrap sample.

Test this function by evaluating the mean of a bootstrap sample associated with the sample generated in Question 1.

3. Use this function in a `for` loop to obtain the sample means of $B = 10000$ bootstrap samples. What is the resulting bootstrap estimate of the mean? What are the bootstrap estimates for the 2.5% and 97.5% quantiles of the distribution of the sample mean? (These are the endpoints of a bootstrap 95 confidence interval for the mean).

4. Show a histogram of the $B = 10000$ sample means. Identify (using `abline`) the population mean, the sample mean of the original sample, the bootstrap estimate of the mean, and the bootstrap 95% confidence interval for the mean.

[1] The content of this worksheet is strongly based on a worksheet designed by Nathalie Vialaneix and Davy Paindaveine.

[2] Recall that the $\Gamma(k,\theta)$ is a continuous distribution admitting the density $f_{k,\theta}(x) = (\Gamma(k))^{-1}\theta^{-k}x^{k-1}\exp(-x/\theta)\mathbb{1}[x > 0]$ where $\Gamma(\cdot)$ is the Euler Gamma function (that is such that $\Gamma(k) = (k-1)!$ for any positive integer k). The mean of the $\Gamma(k,\theta)$ distribution is $k\theta$.

2 The bootstrap using the package boot

Exercise 2. In this exercise, we use the package boot to calculate a bootstrap estimate of the variance of the sample mean. The computational time is compared to the one resulting from a for loop.

1. Using the function rbeta, generate a random sample of size $n = 20$ from the Beta$(1,2)$ distribution[3]. Denoting the sample mean estimator as $\bar{x}$, what is the value of $\bar{x}$ on this particular sample? Compare with the population mean.
2. Compute theoretically the variance of the estimator $\bar{x}$ in the present setup (that is, for a random sample of size 20 from the Beta$(1,2)$ distribution)?
3. Using the sample above and a for loop, obtain a bootstrap estimate for the variance of $\bar{x}$ from $B = 10000$ bootstrap samples. Use the function system.time to show the computational time required to compute the 10000 bootstrap sample means.
4. Load the package boot with the command line

```
library(boot)
```

The function boot requires an argument statistic which is a function of two arguments, namely

- a vector containing the original sample and
- a vector of indices pointing to a subsample on which the statistic (here the mean) will be evaluated (use

```
?boot
```

to see examples of such functions). Create a function

```
boot.mean <- function(a.sample, vector.of.indices) {
  ...
}
```

that returns the sample mean of the bootstrap sample obtained from a.sample by considering observations indexed by vector.of.indices.

5. Use the boot function to generate $B = 10000$ sample means of bootstrap samples. Store the results in an R object named res.boot. Use names(res.boot) to see which information is included in this object. How to access the 10000 sample means (use the help page of the function boot)? Compute the resulting bootstrap estimate of the variance of $\bar{x}$.
6. Compare the computational time required by the boot function with the one in Question 3.

[3] Recall that the Beta(α,β) is a continuous distribution admitting the density $\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}\mathbb{1}[x \in [0,1]]$. Its mean is $\frac{\alpha}{\alpha+\beta}$ and its variance is $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

3 The bootstrap in linear regression models

Exercise 3. This exercise illustrates how the bootstrap can be used in linear regression to obtain

- confidence intervals for the parameters
- and confidence bounds for the prediction.

1. For $\alpha = -1$ and $\beta = 2$, generate a random sample $(X_1, Y_1), \ldots, (X_{20}, Y_{20})$ from the linear regression model

$$Y = \alpha + \beta X + 0.5(\varepsilon - 1)$$

   where X is uniformly distributed over $[0, 1]$, ε is chi-square distributed with 1 degree of freedom, and where X and ε are independent (note that, in this model, the least squares estimators of α and β do *not* follow the usual normal distribution, which makes inference more challenging). Make a scatter plot of this sample. Add to the plot the true linear relationship, namely the straight line $y = 2x - 1$.

2. Use the function `lm` to find the least squares estimates $\hat{\alpha}$ and $\hat{\beta}$. Add to the previous plot the corresponding estimated linear relationship.

3. Use the function `boot` to find a 95% confidence interval for $\hat{\alpha}$ from $B = 5000$ bootstrap samples. Repeat the exercise for $\hat{\beta}$.

4. Use the function `boot` to obtain a 95% confidence interval for each of the predictions associated with $x = 0, 0.01, 0.02, \ldots, 0.99, 1$ (use $B = 5000$ bootstrap samples). Add the corresponding lower and upper bounds for the prediction (as functions of x) to the previous plot.