Bootstrap for quantiles estimation

Author:
Joris Limonier
University of Luxembourg
joris.limonier.001@student.uni.lu

Supervisor:
Nicolas Navet
University of Luxembourg
nicolas.navet@uni.lu

February - May 2021

Contents

List of Figures

1 Introduction

In this project we want to estimate the confidence we have when computing various quantiles. This task may be fairly straightforward when working with known distributions or arbitrary amounts of data but this seldom occurs in engineering or other real world applications. For this reason, here we try to accomplish that task on an unknown distribution and given limited amounts of data.

We define the k-th quantile as follows:

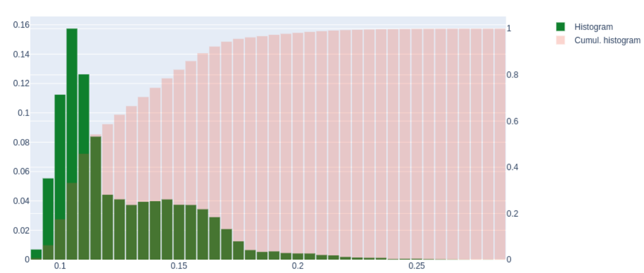
$$q_k := 1 - 10^{-k} \tag{1}$$

which gives $q_1 = 0.9$, $q_2 = 0.99$ ...etc. Where, simply speaking, q_k represents "0" followed by n nines. For our applications, we are mostly interested in q_3, q_4 and q_5.

The data we work with is presented in figure 1a and plotted in the histogram in figure 1b.

Value	Width	Count
0.09	0.005	310
0.095	0.005	2491
0.1	0.005	5058
0.105	0.005	7083
0.11	0.005	5681
0.115	0.005	3771
0.12	0.005	1989
0.125	0.005	1848
0.13	0.005	1676
0.135	0.005	1772
0.14	0.005	1794
0.145	0.005	1846
0.15	0.005	1682
0.155	0.005	1675
0.16	0.005	1544
0.165	0.005	1301
0.17	0.005	936
0.175	0.005	560
0.18	0.005	292
0.185	0.005	235
0.19	0.005	252
0.195	0.005	202
0.2	0.005	188
0.205	0.005	187
0.21	0.005	141
0.215	0.005	129
0.22	0.005	81
0.225	0.005	64
0.23	0.005	56
0.235	0.005	54
0.24	0.005	18
0.245	0.005	24
0.25	0.005	26
0.255	0.005	16
0.26	0.005	11
0.265	0.005	7
0.27	0.005	4
0.275	0.005	1
0.28	0.005	3
0.285	0.005	1

(a) Initial data



(b) Visual representation of the initial data

Figure 1: A look at the initial data

2 Bootstrap

We use the bootstrap to estimate the confidence in the computed quantiles, even with small to moderate sample size.

Let n be the size of the data at our disposal and let q_k be the quantile we want to compute. Here is how we proceed:

1. Do the following R times.
   (a) Draw a random sample of size n with replacement from the initial data.
   (b) Compute q_k on the sample which has just been drawn.
2. Compute the mean over the set of values resulting from step 1b.
3. Compute the confidence intervals (details in section 3)

The plots in figure 2 show the evolution of our estimate for the value of the quantiles as we go through replicates of the bootstraps. The grey areas represent the 95% confidence intervals during that evolution. We will see how to get the confidence intervals from our histogram in section 3.

3 Confidence intervals on the bootstrap

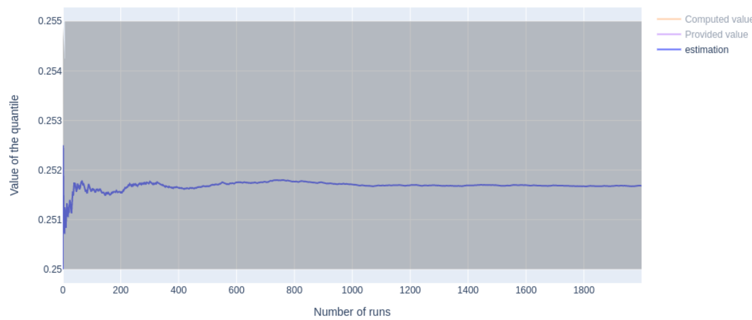
Let us assume that we want to compute the 95% confidence interval.
First, let us note the following:

$$\frac{1-0.95}{2} = 0.025 \qquad \text{and} \qquad 1 - \frac{(1-0.95)}{2} = 0.975 \tag{2}$$

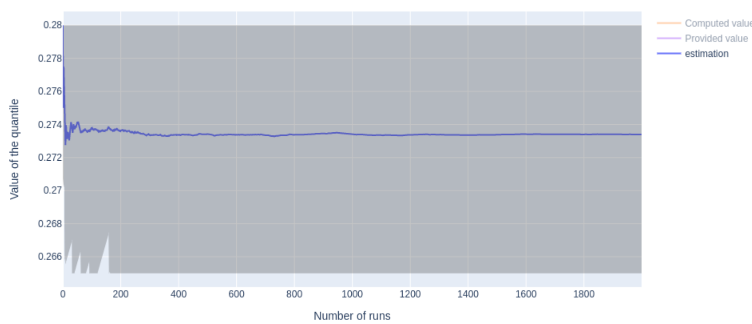
Now we compute the 95% confidence interval. First, we sort the set of all the q's that have been computed for each individual sample. Then we take the 0.025-th and 0.975-th quantile respectively as our lower-bound and upper bound for the confidence intervals.
More generally, for a γ-confidence interval (instead of 95%), one has that the lower and upper bounds of the confidence interval are respectively

$$\gamma_{lo} := \frac{1-\gamma}{2} \qquad \text{and} \qquad \gamma_{up} := 1 - \frac{(1-\gamma)}{2}$$

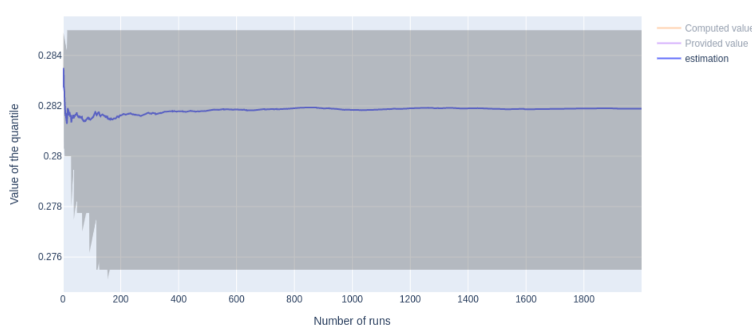
According to our data, it seems that there is close to no change in the estimation of the quantiles after 1000 replications of the bootstrap. The variations are small after 500 replicates already but for safety purposes we consider that we have our final guess after 1000 replicates. As for the confidence intervals, only in some edge cases do we have changes past the 1000 mark.
The fact that our estimate seems to be stable after 1000 replicates does not matter for the number 1000 itself. However, it matters because we seem to "converge" to some value and reach a final value.



(a) Estimation of q_3



(b) Estimation of q_4



(c) Estimation of q_5

Figure 2: Estimation of the quantiles over bootstrap replicates

NB: Our estimate of 1000 replicates is based on empirical evidence. It is not a theoretical result, however, we believe that it is suitable for engineering purposes.

According to the litterature, that is according to Jean-Yves Le Boudec's *Performance Evaluation of Computer and Communication Systems*, there exists some "good value" for the number of bootstrap replicates R. With γ being the confidence level, Le Boudec advises to choose R as follows

$$R = \frac{50}{1-\gamma} - 1 \tag{3}$$

therefore for $\gamma = 95\% = 0.95$ we deduce that the value of R is

$$\begin{aligned} R &= \frac{50}{1-0.95} - 1 \\ &= 1000 - 1 \\ &= 999 \end{aligned}$$

We find that our estimates verifies Le Boudec's result (1000 compared to 999).

A Supplementary tests on other datasets

To verify our approach and make sure that it generalises well, we test it on other datasets.