

# Student job report: Bootstrap high quantiles estimation.

Joris LIMONIER

February - May 2021

## Contents

|          |  |          |
|----------|--|----------|
| <b>1</b> | <b>Introduction</b>                          | <b>2</b> |
| <b>2</b> | <b>Bootstrap</b>                             | <b>3</b> |
| <b>3</b> | <b>Confidence intervals on the bootstrap</b> | <b>3</b> |
| <b>4</b> | <b>Fine tuning</b>                           | <b>5</b> |

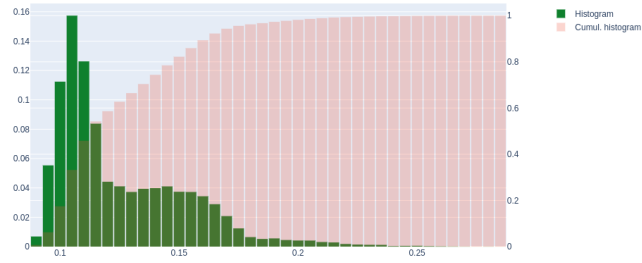
# 1 Introduction

In this project we work on limited data to estimate the values which have a given “low” likelihood. We want to answer questions such as “what is the minimum value which will happen less than 1 case per 10 000 on average ?”. Moreover, we are interested in determining how certain we are of our estimates.

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

We define the  $n$ -th quantile as follows:

$$q_n := 1 - 10^{-n} \quad (1)$$

which gives  $q_1 = 0.9$ ,  $q_2 = 0.99$  ...etc. Where simply speaking  $q_n$  represents “0” followed by  $n$  nines. We are mostly interested in  $q_3$ ,  $q_4$  and  $q_5$ .

## 2 Bootstrap

We use the bootstrap to estimate the value of the quantiles even with small to moderate sample size. That is, we take a random sample of a given length from our data, compute the quantile we are interested in, then repeat the process over multiple runs.

The results from the bootstrap are shown in figure 2:

The plots in figure 2 show the evolution of the quantiles as we go through 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.

**According to our data**, it seems that there is close to no change in the estimation of the quantiles after 1000 repetitions of the bootstrap. The variations are small after 500 runs already but for safety purposes we consider that we have our final guess after 1000 runs. As for the confidence intervals, only in some edge cases do we have changes past the 1000 mark.

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

## 3 Confidence intervals on the bootstrap

Let us fix a quantile that we call  $q$  that we want to estimate. The bootstrap consists of two parts:

- Part 1 Take a random sample from our original data, and determine its value for  $q$ .
- Part 2 Take all the  $q$ 's that have been computed for each individual sample and deduce  $q$  and its confidence interval.

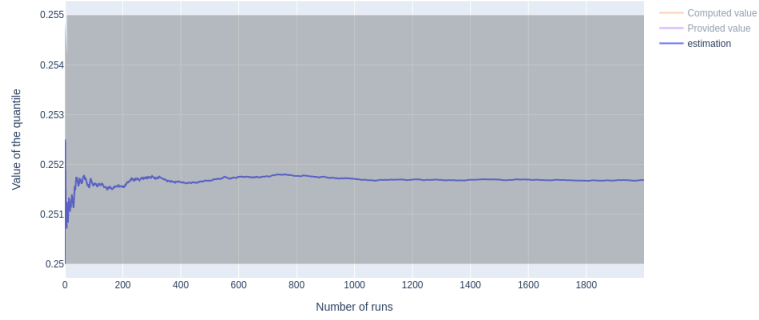
Part 1 is fairly straightforward so we will focus on the second part.

**Estimating the quantile** Our estimation of the quantile  $q$  for the underlying distribution is simply computed by taking the average of all the  $q$ 's of each individual sample.

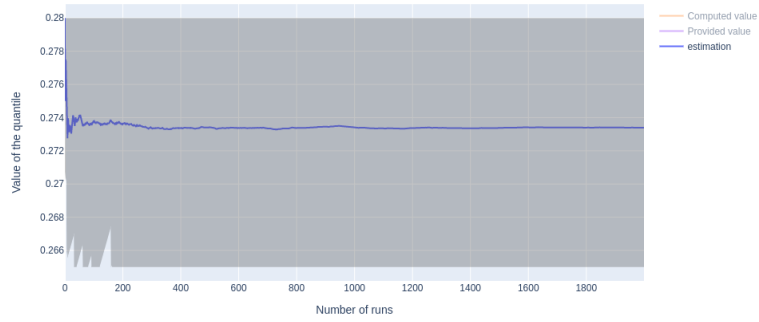
**Computing the confidence intervals** Let's say we want to compute the 95% confidence interval. To do so, we take all the  $q$ 's that have been computed for each individual sample and sort them. Then we take the 0.025-th quantile as our lower bound for the confidence intervals and the 0.975-th quantile as our upper bound for the confidence intervals.

As one may expect, the values 0.025 and 0.975 are found as follows:

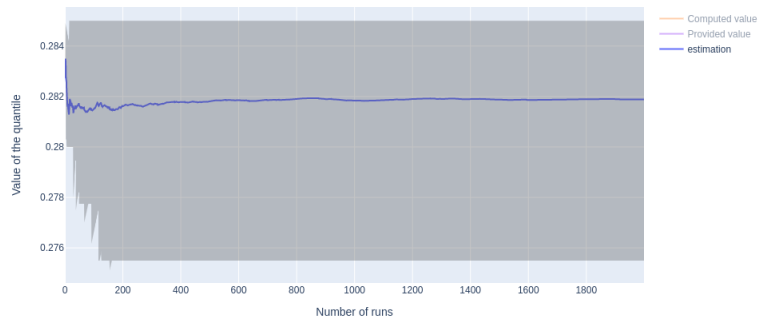
$$\frac{1 - 0.95}{2} = 0.025 \quad \text{and} \quad 1 - \frac{(1 - 0.95)}{2} = 0.975$$



(a) Estimation of  $q_3$



(b) Estimation of  $q_4$



(c) Estimation of  $q_5$

Figure 2: Estimation of the quantiles over bootstrap runs

where the 0.95 comes from the 95% confidence interval.

More generally, for a confidence level of  $\gamma$  (instead of 95%), one has that the lower and upper bounds of the confidence interval are respectively

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

## 4 Fine tuning

Let us assume that we are working with a given data set and we have no way of getting new or more reliable data. We are trying to get the quantiles as precisely as possible and with confidence intervals as small as possible.

There are two parameters of the bootstrap that we can adjust to meet our goals

1. The size of the resamples.
2. The number of resamples.

We touched on the number of resamples already in section 2, however we haven't discussed the size of the resamples.

What impact would increasing the size of the resamples have ? The only time it comes into play is when we take the quantiles of the resamples. Then, what would happen if we took the quantiles of a resample of different size ? Taking the quantile of a smaller resample likely means that we will have less *refinement*. That is, our quantiles at each resample would likely take a smaller set of values. This is further explained by the way we proceed if we have sparse values, which is our case since we work with “high” quantiles (where we have few values available).