Hochschule für Technik
und Wirtschaft Berlin

University of Applied Sciences

Bootstrapping

Term Paper

from

Hagen Wey

Matriculation number: s0575947

Faculty 4 – Informatik, Kommunikation und Wirtschaft –
of the Hochschule für Technik und Wirtschaft Berlin

Term paper in the course "Aktuelle Themen der Informatik"
in the degree programme
Angewandte Informatik

Day of delivery: 16.07.2024

Lecturer: Prof. Dr. Tatiana Ermakova

Inhaltsverzeichnis

Abstract

The objective of this term paper is to provide readers with an introduction to the concepts associated with the bootstrapping process. The following sections aim to provide an overview of the process, as well as an analysis of the advantages and disadvantages of the procedure discussed in this paper. While some basic knowledge of statistics is undoubtedly beneficial for those wishing to gain familiarity with the subject matter, this paper is intended to serve as an introductory guide for any interested reader. Consequently, the fundamental concepts and underlying logic of bootstrapping should be accessible and comprehensible even to those with no prior experience or knowledge in this field. The use of practical examples and illustrative implementations will facilitate a deeper understanding of the subject matter.

Kapitel 1: Introduction

1.1 Background and motivation

In statistical analysis, bootstrapping is a method used for the derivation of robust estimates of standard errors and confidence intervals. This is applied to estimates such as mean, median, proportion, odds ratio, correlation coefficient or regression coefficient. Developed by Bradley Efron towards the end of the 1970s in "Bootstrap methods: another look at the jackknife"(1979), inspired by the jackknife technique, the bootstrap remains one of the most significant approaches in contemporary statistics, particularly as an alternative to parametric estimates. The flexibility of this method in addressing uncertainty in estimates, particularly in the context of smaller or non-normally distributed samples, has made it a valuable tool in a range of business domains, including biostatistics, financial analysis and machine learning. This particular area is of outstanding importance for our course. In this context, the topic of bootstrapping exemplifies the relationship between theoretical concepts and their practical application. This will be demonstrated later in the course through the implementation of clear, yet uninterpretable, results.

1.2 Objectives of the paper

The objective of this paper is to provide the reader with an insight into the concept of bootstrapping. To this end, we will first elucidate some pivotal terminology and fundamental mathematical principles, with a view to fostering a more nuanced comprehension. Thereafter, we will examine the structural underpinnings of bootstrapping and identify the key elements that warrant consideration prior to reinvigorating the process through the use of a theoretical exemplar. Finally, we will explore a potential implementation and utilize this to elucidate the individual steps in what we term "plots".

Kapitel 2: Mathematical/statistical foundations of bootstrapping

2.1 Definition of terms

2.1.1 Standard error

The standard error represents a significant statistical indicator, measuring the dispersion of sample mean values in relation to the true population value. It indicates the degree of accuracy with which the sample mean approximates the true population mean. A smaller standard error value indicates a more accurate estimate of the population mean by the sample mean. Conversely, a larger standard error value indicates an inaccurate estimate. The formula for the standard error is as follows:

$$SE = \frac{s}{\sqrt{n}}$$

where s is the standard deviation of the sample and n is the sample size.

In the context of bootstrapping, the standard error is employed for the purpose of calculating confidence intervals for estimates. The aforementioned intervals indicate a range within which the true population parameter is likely to be situated with a high degree of probability.

2.1.2 Confidence intervalls

Confidence intervals represent a fundamental concept in the field of statistics and constitute a crucial element of the bootstrap methodology. A confidence interval, or CI for short, represents the range within which a parameter (e.g. the mean value) is likely to fall with a certain probability. It is common practice among analysts to utilise confidence intervals that encompass either 95

The calculation of a confidence interval for the mean is typically performed as follows:

$$\bar{x} \pm z^* \cdot SE$$

where $\bar{x}$ is the sample mean, z^* is the critical value from the standard normal distribution for the desired confidence level (e.g. 1.96 for a 95% confidence interval) and SE is the standard error.

2.2 Bootstrapping methodology

Having established the two most crucial elements of bootstrapping and the means of calculating them, we may now proceed to the subject matter itself: bootstrapping.

2.2.1 Theory

The fundamental concept of bootstrapping is the utilisation of sample data to infer conclusions regarding an estimated value (e.g. the sample mean) for a population parameter θ (e.g. the population mean). Consequently, bootstrapping represents a methodology of replicate sampling. In this process, random samples are drawn independently of each other from existing sample data with the same sample size n, from which conclusions can then be drawn. This process enables the generation of empirical distributions and the calculation of confidence intervals.

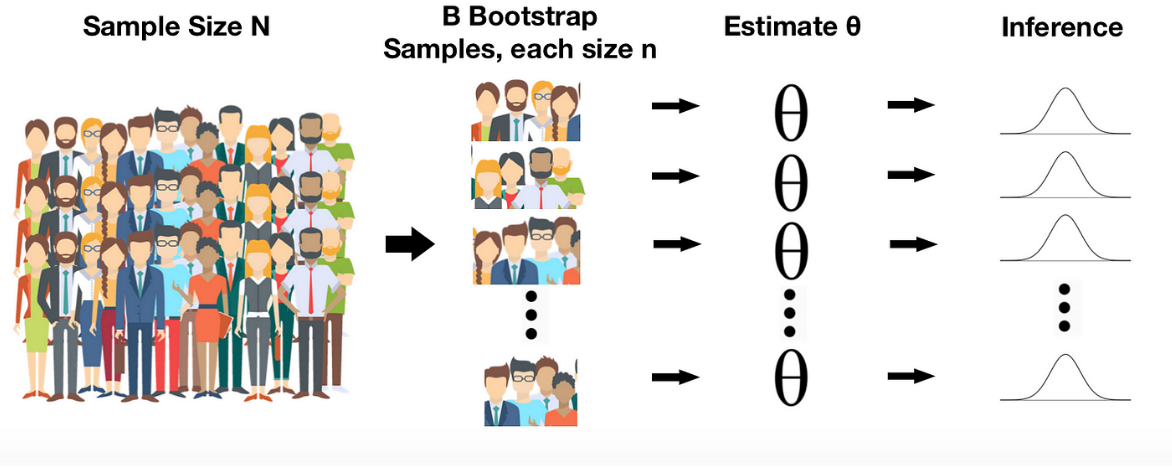

Abbildung 2.1: Bootstrapping example

Figure 2.1 provides a further illustration of the process in an abstract form. The accumulated knowledge, as illustrated in Figure 2.1, gives rise to the following process:

1. **Drawing the bootstrap samples**:
   - Assume we have a sample of n data points $X = \{x_1, x_2, \ldots, x_n\}$.
   - Draw B bootstrap samples from X with replacement. Each bootstrap sample has the same size as the original sample n.

2. **Calculation of bootstraps**:
    - For each of the B bootstrap samples, compute the desired statistic $\hat{\theta}^*$ (e.g., mean, median, standard deviation).
3. **Estimation of the confidence interval**:
    - Determine the $(100 \cdot (1 - \alpha/2))$-th percentiles of the distribution of $\hat{\theta}^*$ to obtain a $(1 - \alpha)$-confidence interval, where α is the significance level.

2.2.2 Example

In order to provide a more illustrative example of the bootstrapping process, we will examine a fictitious case study from a practical context. In the field of medical research, bootstrapping represents a valuable methodology for estimating the standard error and calculating confidence intervals for a range of parameters. It is therefore appropriate to integrate this example into the medical problems sector, given that it is a topic that we always address in this course.

Let us consider a hypothetical scenario in which a research team is investigating the average efficacy of a novel pharmaceutical agent for the reduction of blood pressure in patients diagnosed with hypertension.

Steps of Bootstrapping:

1. **Data collection**:
    - A clinical trial was conducted by the research team, who collected a sample of 80 patients with hypertension who received the new drug.
2. **Application of the bootstrapping method**:
    - **The bootstrap samples are drawn as follows:** A total of 1,000 bootstrap samples are drawn from the 80 available patient data sets.
    - **Calculation of the bootstrap statistic:** The mean reduction in blood pressure following treatment is determined for each of the 1,000 bootstrap samples.
    - **The estimation of the confidence interval is as follows:** The 95% confidence interval for the mean blood pressure drop is determined by employing the 2.5% and 97.5% percentiles of the bootstrap mean values.
3. **Interpretation of the results**:
    - The research team obtained a 95% confidence interval for the average drop in blood pressure, for example, from 15 to 10 mmHg. This indicates that the actual mean reduction in blood pressure following treatment falls within this interval with 95% confidence.

The utilisation of bootstrapping methodology allows the research team to make a well-founded assertion regarding the efficacy of the recently investigated pharmacological agent. This may subsequently result in the drug being approved or a similar outcome.

Kapitel 3: Implementation of bootstrapping

3.1 Algorithm and pseudocode

3.2 Implementation in Python

Kapitel 4: Conclusion

Kapitel 5: References

https://chrisbogner.github.io/datenanalyse/bootstrapping-und-konfidenzintervalle.html https://de.wikipedia.org/wiki/Konfidenzintervall https://datatab.de/tutorial/konfidenzintervall https://www.investopedia.com/terms/c/confidenceinterval.asp https://de.wikipedia.org/wiki/Standardfel https://dept.stat.lsa.umich.edu/ kshedden/introds/ https://www.ibm.com/docs/de/spss-statistics/saas?topic=bootstrapping- https://towardsdatascience.com/an-introduction-to-the-bootstrap-method-58bcb51b4d60

Link to Latex Template: https://github.com/tscheffl/HTW-Thesis

Quelltextverzeichnis