Technical University Dortmund

Monte Carlo Simulations: Theory and Practice

Project 5: Bootstrapping

Lecturers:

Prof. Dr. Katja Ickstadt

M. Sc. Zeyu Ding

Group number: 05

Authors:

Mohammad Istiak Mahbub

Naim Ahmad

August 12, 2024

Contents

1 Introduction

Bootstrapping is a straightforward and powerful statistical method that helps to assess the accuracy of the estimates, especially when we don't know the true distribution of the data. The technique involves repeatedly sampling from the original data, with replacement, to create new samples. By analyzing these samples, we can predict and calculate measures such as standard errors, confidence intervals, and perform hypothesis tests, which indicate the reliability of the estimates. This method is particularly useful when working with small datasets or when information about the overall population is lacking. Bootstrapping allows researchers to make reliable inferences without depending on complex mathematical assumptions, making it an accessible and practical tool for various types of data analysis (Bittmann, 2021).

The primary goal of this project is to explore the application of the bootstrapping method in statistical analysis, specifically focusing on its use in estimating the mean, calculating standard errors, constructing various types of confidence intervals, and performing hypothesis tests. The content of the project includes an in-depth discussion of the bootstrapping technique, practical examples to illustrate its use, and a critical analysis of its effectiveness. By leveraging bootstrapping, the project aims to demonstrate how reliable statistical inferences can be made even when traditional parametric assumptions are not met.

The approach to solving the problem involves a structured methodology that begins with the generation of bootstrap samples from the original dataset. Each bootstrap sample is used to calculate the mean, which serves as the statistic of interest. Following this, standard errors are calculated, and confidence intervals are constructed using various methods, including Normal, Percentile, Bias Corrected (BC), and Bias Corrected and Accelerated (BCa) intervals. After constructing these confidence intervals, the methodology proceeds with hypothesis testing to further evaluate the statistical significance of the estimates.

The report is organized into several sections. Following this introduction, Section 2 outlines the key statistical challenges the project addresses and the limitations of traditional methods that bootstrapping seeks to overcome. Section 3 details the bootstrapping technique, including the mathematical foundations and the statistical measures estimated, such as standard errors, confidence intervals, and hypothesis tests. Section 4 covers the application of bootstrapping to simulated datasets, including hypothesis testing and the evaluation of the method's robustness. Finally, Section 5 concludes the report by summarizing the findings and suggesting potential directions for future research.

2 Problem Definition

2.1 Overview of bootstrapping and methodology

In statistical analysis, accurately estimating population parameters, such as the mean, and assessing the reliability of these estimates is crucial, particularly when dealing with small sample sizes or when the underlying distribution of the data is unknown. Traditional parametric methods often rely on assumptions that may not hold in real-world data, leading to potential inaccuracies in inference. This project seeks to address the challenge of making reliable statistical inferences without relying on these strict assumptions by employing the bootstrapping method. Specifically, the project focuses on using bootstrapping to estimate the mean, calculate standard errors, construct various types of confidence intervals (Normal, Percentile, Bias Corrected, and Bias Corrected and Accelerated), and perform hypothesis tests. The goal is to demonstrate how bootstrapping can provide robust estimates and inferences in situations where traditional methods might fail or be inadequate. To explore the fundamental principles and processes of bootstrapping, we initially utilized synthetic data, allowing us to test and refine our methodology in a controlled environment.

2.2 Problem statement and hypothesis

The objective of this report is to estimate the average score of students in a class at TU Dortmund University, where the total student population is 100. Due to constraints that prevent surveying all students individually, we employ the bootstrapping technique as an alternative method for estimation. To simulate the data collection process, we have used synthetic data representing the scores of 17 randomly selected students from the class. From this initial sample, we generate numerous bootstrap samples by repeatedly resampling with replacement, each time selecting 17 scores. For each bootstrap sample, we calculate the mean score, and by repeating this process 1000 times, we obtain a distribution of 1000 mean estimates. Using these estimates, we construct confidence intervals to quantify the uncertainty surrounding the average score of the entire class.

To test our approach, we formulate the following hypothesis: *The average score of students in the class is equal to a hypothesized value μ_0 (e.g., the university's expected average score for the course).* The report aims to demonstrate the effectiveness of bootstrapping in estimating population parameters and constructing confidence intervals when a full census is not feasible. Through this process, we can determine with 95% confidence whether the true average score lies within the constructed interval, thereby providing a robust statistical basis for decision-making.

2.3 Application to real-world data

In addition to synthetic data, we applied our bootstrapping approach to a real-world dataset, specifically the 'Election Forecast Based on 2000 Surveys' (Wikipedia contributors, 2024), to predict confidence intervals for vote percentages. Extensive Monte Carlo simulations were conducted to ensure the robustness of our bootstrap estimates, providing deeper insights into the variability and reliability of the statistical measures obtained. By comparing the predicted confidence intervals with actual historical outcomes, we demonstrated the practical applicability and accuracy of the bootstrapping method in real-world scenarios. This comprehensive approach underscores the versatility and effectiveness of bootstrapping as a powerful tool in both theoretical and applied statistical analysis.

3 Statistical Methods

3.1 Basic statistical concepts

3.1.1 Arithmetic mean

The arithmetic mean (or mean) is the sum of all values divided by the total number of values. The average is another term for the mean (Moore et al., 2007).

$$mean, \bar{x} = \frac{x_1 + x_2 + ... + x_n}{n}$$

$$mean, \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

3.1.2 Standard deviation

Standard deviation, denoted by σ, is a fundamental statistical measure used to quantify the dispersion or spread of a set of data points relative to the mean. It essentially reflects how much the individual data points in a dataset deviate from the average value, offering a numerical summary of variability within the dataset.

Mathematically, the standard deviation is calculated using the formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

where, x_i represents each individual observation, $\bar{x}$ is the mean of all observations, and n denotes the total number of observations in the dataset.

This measure is particularly useful because it maintains the same units as the original data, making it easier to interpret the extent of variability. A smaller standard deviation suggests that the data points are closely clustered around the mean, indicating low variability. Conversely, a larger standard deviation indicates that the data points are more spread out, reflecting higher variability within the dataset.(Bittmann, 2021)

3.1.3 Standard error

The standard error (SE) is often confused with the standard deviation, but it serves a distinct purpose in statistics. While the standard deviation measures the spread of individual data points within a sample around the mean, the standard error is used to describe the accuracy of a sample statistic as an estimate of the population parameter. Specifically, the standard error quantifies the variability of the sample mean by considering how different the mean of various samples from the same population could be.

Mathematically, the standard error of the mean (SEM) is calculated using the formula:

$$\text{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the sample, and n is the sample size.

The standard error is crucial in inferential statistics because it helps us understand how well our sample mean estimates the true population mean. The SEM decreases as the sample size increases, reflecting the fact that larger samples provide more reliable estimates of the population parameter. Essentially, the smaller the SEM, the closer the sample mean is likely to be to the actual population mean, reducing uncertainty in our estimates.(Bittmann, 2021)

3.1.4 Alpha (α) and p-value

Alpha (α), also known as the significance level, is a predetermined threshold set before conducting a hypothesis test. It signifies the probability of rejecting the null hypothesis when it is true, representing the likelihood of a Type I error. The p-value, on the other hand, quantifies the probability of observing a test statistic at least as extreme as the observed data, under the assumption that the null hypothesis is true. If the p-value is less than or equal to the alpha level, typically set at 0.05, there is strong evidence against the null hypothesis, leading to its rejection. However, it is important to remember that 0.05 is an arbitrary choice for the alpha level and has been widely criticized for being too high. Consequently, a small p-value (≤ 0.05) suggests significant evidence against the null hypothesis. In summary, when the p-value is less than the chosen alpha level, the null hypothesis is rejected, with an alpha-level chance of committing a Type I error, indicating that the null hypothesis is rejected when it is actually true (Cleophas et al., 2006) (Forero, 2023).

3.2 Bootstrapping

The fundamental idea behind bootstrapping is to treat the available data sample as a proxy for the entire population. By repeatedly resampling the data, we generate many new datasets (bootstrap samples) that are similar to the original sample. This approach allows us to estimate the sampling distribution of a statistic without making strong parametric assumptions.

The core idea of bootstrapping is to use these multiple samples to create an empirical distribution of the statistic of interest, which can be a mean, median, standard deviation, or any other statistic. This empirical distribution allows us to estimate properties such as the standard error, confidence intervals, and bias of the statistic without the need for complex analytical formulas.

The bootstrapping process can be summarized in the following steps:

1. **Resampling:** From the original dataset containing n observations, generate B bootstrap samples. Each bootstrap sample is created by randomly drawing n observations from the original dataset with replacement.
2. **Statistic calculation:** For each of the B bootstrap samples, calculate the statistic of interest (e.g., mean, median). This will yield B estimates of the statistic, denoted as $\hat{\theta}^{*1}, \hat{\theta}^{*2}, \ldots, \hat{\theta}^{*B}$.
3. **Estimate the distribution:** Using the B bootstrap estimates, construct the empirical distribution of the statistic. The standard deviation of this distribution provides an estimate of the standard error of the statistic(Bittmann, 2021) (Robert and Casella, 2004).

3.3 Confidence interval

In this report, we have utilized four different techniques for constructing confidence intervals: normal intervals, percentile intervals, bias-corrected (BC) intervals, and bias-corrected and accelerated (BCa) intervals.

3.3.1 Normal intervals

A **normal interval** is a method used to calculate confidence intervals by assuming that the sampling distribution of the estimator follows a normal distribution. In bootstrapping, this approach involves calculating the bootstrapped standard error of the statistic of interest and then using this to generate symmetrical confidence intervals around the point estimate.

The general formula for a 95% normal confidence interval is:

$$\mathrm{CI}_{\theta}^{95\%} = \theta \pm 1.96 \cdot \mathrm{SE}_{\theta}$$

where:

- θ is the point estimate of the parameter (e.g., the sample mean).
- 1.96 is the critical value from the standard normal distribution for a 95% confidence level.
- SE_θ is the bootstrapped standard error of the estimate θ.

For this interval to be valid, it is essential that the sampling distribution of the estimator is approximately normal. However, this assumption might not always hold, especially in cases where the data is skewed or where the sample size is small.

One of the limitations of the normal interval approach is that it may produce questionable results if the underlying assumptions are violated. For example, if the standard error is large, the confidence interval might include impossible values (such as negative values for a variable that should only be positive). This highlights the importance of verifying the normality assumption when using this method to ensure that the resulting confidence intervals are meaningful (Bittmann, 2021).

3.3.2 Percentile intervals

Percentile intervals offer an alternative method to calculating confidence intervals by utilizing the entire bootstrapped distribution rather than relying solely on the bootstrapped standard error. This approach directly visualizes the overall distribution of the bootstrap samples, providing a more intuitive understanding of the confidence intervals.

The key idea behind percentile intervals is to generate a large number of bootstrapped samples and then sort them. By removing the most extreme values from both ends of the sorted list, we can identify the bounds of the confidence interval. For example, to calculate a 95% confidence interval, we would exclude the lowest 2.5% and the highest 2.5% of the bootstrap samples.

Mathematically, the lower and upper bounds of the α percentile confidence interval can be expressed as:

$$\text{CI}_l^\alpha = \text{Percentile}\left(\frac{\alpha}{2} \cdot 100\right)$$

$$\text{CI}_u^\alpha = \text{Percentile}\left(1 - \frac{\alpha}{2}\right) \cdot 100$$

where:

- α represents the desired confidence level (e.g., 0.05 for a 95% confidence interval).
- Percentile indicates the corresponding percentile rank in the sorted bootstrap distribution.

This method is particularly useful when the distribution of the bootstrap samples is not symmetric, as it adjusts the confidence interval to fit the actual shape of the distribution. By focusing on the actual bootstrap results, percentile intervals avoid some of the assumptions required by normal intervals, making them a flexible and robust choice for confidence interval estimation (Bittmann, 2021).

3.3.3 Bias-corrected intervals (BC)

Bias-corrected intervals (BC) are an extension of percentile intervals that aim to adjust for bias in the bootstrap distribution. The bias correction accounts for the systematic deviation between the bootstrap estimate and the original sample estimate.

The bias correction factor, z_0, is calculated as follows:

$$z_0 = Z^{-1}(P_{\text{Boot}})$$

where:

- Z^{-1} is the inverse cumulative standard normal function.
- P_{Boot} represents the proportion of bootstrap estimates that are less than or equal to the original sample estimate $\hat{\theta}$.

The corrected lower and upper bounds of the confidence interval are then adjusted using this bias correction factor:

$$Y_{95\%L} = z_0 + Z^{-1}\left(\frac{\alpha}{2}\right)$$

$$Y_{95\%U} = z_0 + Z^{-1}\left(1 - \frac{\alpha}{2}\right)$$

where:

- α is the significance level (e.g., 0.05 for a 95% confidence interval).
- Z^{-1} is used again to adjust the bounds based on the bias correction factor.

This correction shifts the confidence interval to account for any skewness or bias in the bootstrap distribution, leading to more accurate interval estimates (Bittmann, 2021).

3.3.4 Bias-corrected and accelerated intervals (BCa)

Bias-Corrected and Accelerated Intervals (BCa) extend the concept of bias-corrected intervals by incorporating an additional correction known as acceleration. This correction accounts for the rate of change of the bootstrap standard error concerning the statistic being estimated, thereby improving the accuracy of the confidence intervals.

The BCa interval formula adjusts both for bias and acceleration and is given by:

$$Y_{95\%L} = \frac{z_0 + Z^{-1}\left(\frac{\alpha}{2}\right)}{1 - a\left(z_0 + Z^{-1}\left(\frac{\alpha}{2}\right)\right)}$$

$$Y_{95\%U} = \frac{z_0 + Z^{-1}\left(1 - \frac{\alpha}{2}\right)}{1 - a\left(z_0 + Z^{-1}\left(1 - \frac{\alpha}{2}\right)\right)}$$

where:

- z_0 is the bias correction factor, calculated as $z_0 = Z^{-1}(P_{\text{Boot}})$.
- a is the acceleration constant, calculated using jackknife resampling.
- Z^{-1} is the inverse cumulative standard normal function.
- α is the significance level (e.g., 0.05 for a 95% confidence interval).

The acceleration constant a is determined using the jackknife method, which resamples the data by leaving out one observation at a time and calculates the variability of the estimator (Bittmann, 2021).

3.4 Hypothesis test

A hypothesis test is a statistical method used to make decisions or inferences about a population parameter based on a sample of data (Bittmann, 2021).

Formulation of Hypotheses:

- **Null hypothesis (H_0):** This is a statement that there is no effect or no difference, and it represents the status quo or a condition to be tested. It's typically formulated in a conservative manner. For example, if we want to test whether the mean score of a population is equal to a specific value, the null hypothesis could be:

$$H_0 : \mu = \mu_0$$

- **Alternative hypothesis (H_1):** This is the statement that there is an effect or a difference, and it represents the outcome that we are trying to find evidence for. The alternative hypothesis is the complement of the null hypothesis. It could be:

$$H_1 : \mu \neq \mu_0 \quad \text{(two-tailed)}$$

or

$$H_1 : \mu > \mu_0 \quad \text{(one-tailed, if testing for an increase)}$$

or

$$H_1 : \mu < \mu_0 \quad \text{(one-tailed, if testing for a decrease)}$$

3.5 Software and tools

The analysis in this report was conducted using the R statistical software version 4.2.2 (R Core Team, 2023) and a combination of base R functions and additional packages. Key functions included `mean()`, `sd()`, `sample()`, and `quantile()` for statistical computations, and `plot()`, `hist()`, and `legend()` for data visualization. The `"boot"` package (Canty and Ripley, 2021) was also employed to compare the custom bootstrap method against R's built-in bootstrap function. This setup allowed for efficient execution of bootstrap resampling, hypothesis testing, and visualization, ensuring accurate and reproducible results throughout the analysis.

3.6 Underlying assumptions and justification

The validity of bootstrapping hinges on several key assumptions about the data. Most importantly, it assumes that the sample data is representative of the population from which it is drawn and that the data points are independent and identically distributed (i.i.d.). These assumptions are crucial because they ensure that the resampling process accurately captures the underlying variability within the population. When these assumptions are satisfied, bootstrapping can provide reliable estimates for standard errors, confidence intervals, and hypothesis tests without relying on strict parametric models. However, if the assumptions are violated, the reliability of the results may be compromised, potentially leading to skewed or inaccurate statistical inferences (Field, 2009) (Forero, 2023).

In this experiment, the i.i.d. assumption is justified by the nature of our data collection process. The data was collected through a randomized sampling method, ensuring that each observation is independent of the others. This randomization process helps prevent any underlying relationships or dependencies between the data points. Additionally, all observations are drawn from the same population and are therefore assumed to follow the same underlying distribution. As a result, the i.i.d. assumption holds true in this context.

4 Statistical Analysis

In this section, we present a comprehensive statistical analysis of the data, focusing on the application of bootstrapping techniques to estimate various statistical measures. The analysis includes summary statistics, estimation of standard errors, construction of confidence intervals, and hypothesis testing. Additionally, we assess bias and variance, explore computational complexity, and evaluate the convergence of bootstrap means. Each analysis step is meticulously documented, with results presented in a clear and structured manner, rounded to two decimal places for consistency. Each result is interpreted with regard to the problem definition, ensuring a thorough understanding of the statistical implications.

4.1 Summary statistics of the dataset

First, we present the summary statistics of the dataset used in the analysis. The dataset consists of the following values:

- **Dataset:** $\{18, 28, 30, 31, 35, 37, 38, 46, 50, 53, 54, 59, 61, 63, 67, 69, 71\}$
- **Mean:** 47.64
- **Standard Deviation:** 16.23
- **Median:** 50

The dataset includes scores from a sample of 17 students. The mean score is 47.64, indicating the average score across the dataset. The standard deviation is 16.23, reflecting the variability or dispersion of the scores from the mean. The median score, which is the middle value when the data is ordered, is 50.

4.2 Standard error estimation using bootstrap

To estimate the standard error of the mean for the dataset, we employed a custom bootstrap method specifically designed for this analysis. This custom method utilizes vectorization to efficiently generate bootstrap samples and compute the desired statistic.

The process involved generating 1000 bootstrap samples by resampling with replacement from the original dataset. For each bootstrap sample, the mean was calculated, and the distribution of these bootstrap means was used to estimate the standard error.

- **Estimated Mean from Bootstrap Samples**: The estimated mean calculated from the bootstrap samples is 47.65. - **Standard Error of the Mean**: The standard error of the mean, calculated using the distribution of bootstrap means, is 3.89.

Note: The full implementation of the custom bootstrap method is provided in Appendix A: Pseudocode of custom bootstrap method.

4.3 Confidence interval estimation

In this analysis, we employed the bootstrapping technique to estimate the mean and construct various types of confidence intervals (CIs) for the dataset. By generating 1000 bootstrap samples, we calculated the mean for each sample and used the distribution of these bootstrap means to construct the following confidence intervals:

4.3.1 Normal confidence interval

The Normal Confidence Interval assumes that the distribution of the bootstrap means follows a normal distribution. Based on the standard error calculated from the bootstrap means, the 95% Normal Confidence Interval is:

40.03 to 55.27

This interval is symmetric around the mean and reflects the typical assumption of normality.

4.3.2 Percentile confidence interval

The percentile confidence interval is constructed by directly using the percentiles of the bootstrap distribution. The 95% Percentile Confidence Interval is:

40.17 to 54.71

This interval is derived purely from the empirical distribution of the bootstrap means, without relying on normality assumptions.

4.3.3 Bias-corrected (BC) confidence interval

The bias-corrected confidence interval adjusts the percentile method by accounting for any bias present in the bootstrap estimates. The 95% Bias-Corrected Confidence Interval is:

39.91 to 55.14

This interval corrects for the potential bias in the bootstrap sampling, providing a more accurate estimate than the percentile method alone.

4.3.4 Bias-corrected and accelerated (BCa) confidence interval

The bias-corrected and accelerated confidence interval further adjusts for both bias and skewness in the bootstrap distribution. The 95% Bias-Corrected and Accelerated Confidence Interval is:

36.77 to 59.09

This interval is wider than the other intervals, reflecting the additional adjustments for bias and acceleration, which accounts for the asymmetry in the bootstrap distribution.

4.3.5 Comparison and interpretation

The results of the confidence intervals show varying ranges depending on the method used. The normal and percentile intervals are relatively close, while the bias-corrected and BCa intervals provide a broader range, indicating the presence of bias and skewness in the bootstrap distribution. The BCa interval, being the most conservative, suggests that the true mean could be as low as 36.77 or as high as 59.09 with 95% confidence. These results underscore the importance of selecting an appropriate method for confidence interval estimation, particularly when the underlying data distribution is unknown or may deviate from normality.

4.4 Bias and variance analysis

In addition to constructing confidence intervals, we conducted a bias and variance analysis of the bootstrap estimates to further assess the accuracy and reliability of our estimator.

4.4.1 Bias estimation

Bias is defined as the difference between the expected value of the estimator (the average of the bootstrap means) and the true parameter value (the sample mean). It provides insight into whether the estimator systematically overestimates or underestimates the true parameter. In this analysis, the bias of the bootstrapped mean was calculated as:

Bias of the bootstrap estimates: 0.0054

The bias estimate is close to zero, indicating that the bootstrapped mean is an unbiased estimator of the true sample mean, with minimal systematic error.

4.4.2 Variance estimation

Variance measures the variability of the estimator, indicating how much the bootstrap estimates fluctuate around their mean. A lower variance indicates that the estimates are more stable, while a higher variance suggests greater variability. The variance of the bootstrapped mean was calculated as:

Variance of the bootstrap estimates: 15.11

This variance estimate reflects the spread of the bootstrap means and provides an indication of the reliability of the estimator. A relatively low variance implies that the bootstrapped mean is consistent across different resamples.

4.4.3 Analysis and interpretation

The bias estimate being near zero suggests that the bootstrap method used in this analysis provides an accurate estimation of the mean, with little to no systematic error. The variance estimate indicates the degree of variability in the bootstrap estimates; a variance of 15.11 suggests that while the estimates are fairly consistent, there is still some spread around the mean.

Overall, the bias and variance analysis confirms that the bootstrap method is both accurate and reliable, making it a suitable approach for estimating the mean and constructing confidence intervals, particularly in cases where the underlying distribution of the data is unknown.

4.5 Hypothesis testing

In this subsection, we assess whether the average score of students in the class is significantly different from a hypothesized mean value, μ_0. The hypothesized mean value, in this case, is set to 50, which could represent the university's expected average score for the course. To test this hypothesis, we employed the bootstrap method to generate a distribution of sample means and used it to perform a two-tailed hypothesis test.

4.5.1 Null and alternative hypotheses

The hypotheses are formulated as follows:

- **Null hypothesis (**H_0**)**: The average score of students is equal to the hypothesized value μ_0.

$$H_0 : \mu = \mu_0$$

- **Alternative hypothesis (**H_1**)**: The average score of students is not equal to the hypothesized value μ_0.

$$H_1 : \mu \neq \mu_0$$

4.5.2 Methodology

To conduct the hypothesis test, we followed these steps:

1. **Bootstrap sampling**: We generated 1000 bootstrap samples from the original dataset, each time calculating the mean of the sample.
2. **P-value calculation**: The p-value was computed as the proportion of bootstrap means where the absolute difference from the observed mean was greater than or equal to the absolute difference from μ_0. This approach is consistent with the requirements of a two-tailed test, where we are interested in detecting any significant difference from the hypothesized mean, regardless of direction.
3. **Decision rule**: We set the significance level (α) at 0.05. If the computed p-value is less than α, we reject the null hypothesis in favor of the alternative. Otherwise, we fail to reject the null hypothesis.

4.5.3 Results

The p-value obtained from the hypothesis test was 0.575. Since the p-value is greater than the significance level ($\alpha = 0.05$), we **fail to reject the null hypothesis**. This result indicates that there is no significant difference between the observed average score and the hypothesized value of 50.

4.5.4 Conclusion

Based on the results of the two-tailed hypothesis test, we conclude that the average score of students in the class does not differ significantly from the hypothesized mean of 50.

This finding suggests that the observed data is consistent with the assumption that the average score is 50, and there is insufficient evidence to suggest a difference. The use of the bootstrap method has provided a robust and non-parametric approach to testing this hypothesis, reinforcing the validity of the results obtained.

4.6 Computational complexity and execution time

This section explores the computational complexity of the bootstrapping process by analyzing the relationship between the number of bootstrap samples (B) and execution time. Figure 1 presents a log-log plot that illustrates this relationship, with execution time (in seconds) plotted against the number of bootstrap samples. The plot reveals a strong linear trend, with an R-squared value of 0.99, indicating a very strong positive correlation between the number of bootstrap samples and execution time. The slope of the regression line is 0.78, suggesting that execution time increases sub-linearly as the number of bootstrap samples

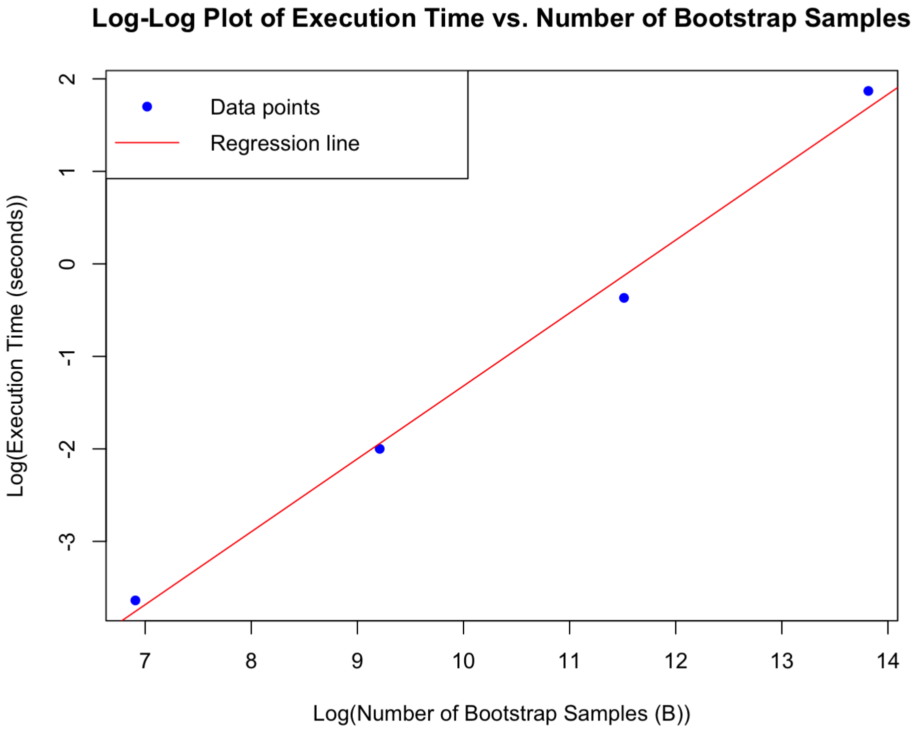

Figure 1: Log-Log plot of execution time vs. the number of bootstrap samples.

grows. This means that while execution time does increase, it does so at a slower rate than the increase in the number of samples.

These findings highlight the efficiency and scalability of the bootstrapping process, making it suitable for large-scale applications where computational resources need to be optimized. The strong predictability of the execution time based on the number of bootstrap samples allows for better planning in resource-intensive scenarios.

4.7 Convergence of bootstrap mean

To analyze the convergence of the bootstrap mean to the true mean as the number of bootstrap samples increases. Figure 2 shows the cumulative bootstrap mean (blue line) against the number of bootstrap samples (B), with the true mean (red dashed line) at 47.64. Initially, the bootstrap mean fluctuates, but as more samples are added, it converges towards the true mean. This convergence illustrates the effectiveness of bootstrapping, showing

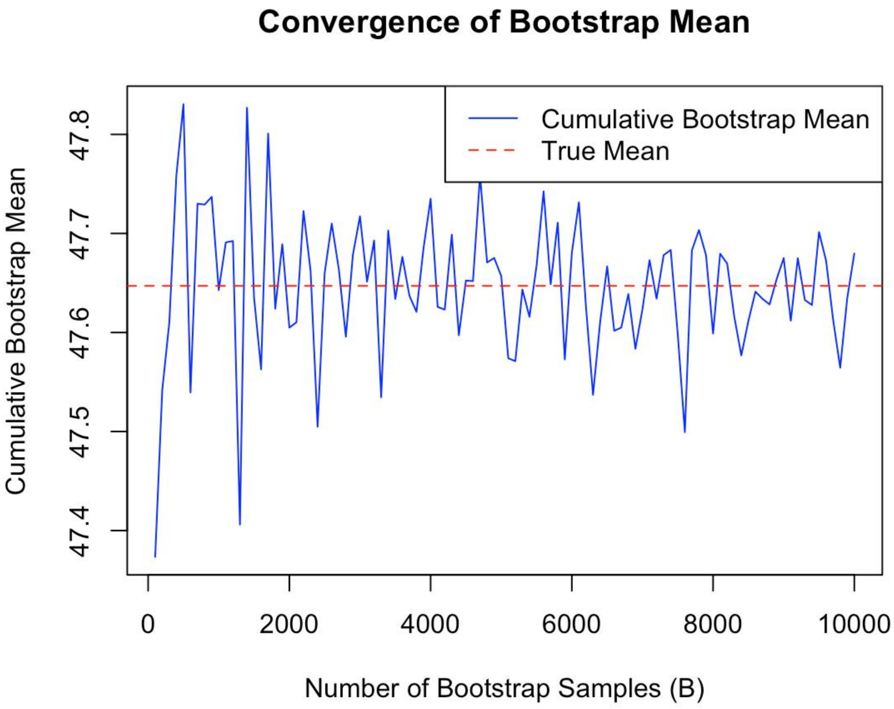

Figure 2: Convergence of bootstrap mean to true mean.

that as the number of bootstrap samples increases, the estimate becomes more accurate and stabilizes around the true mean. The figure underscores bootstrapping's reliability for accurate parameter estimation, even when direct computation is challenging.

In summary, figure 2 demonstrates how bootstrapping produces increasingly precise estimates as more samples are used, confirming its robustness as a statistical tool.

4.8 Histogram analysis of original and bootstrap sample data

To gain further insights into the distributional characteristics of the data, we performed a histogram analysis comparing the original dataset with the bootstrap samples. The histograms are shown in Figure 3.

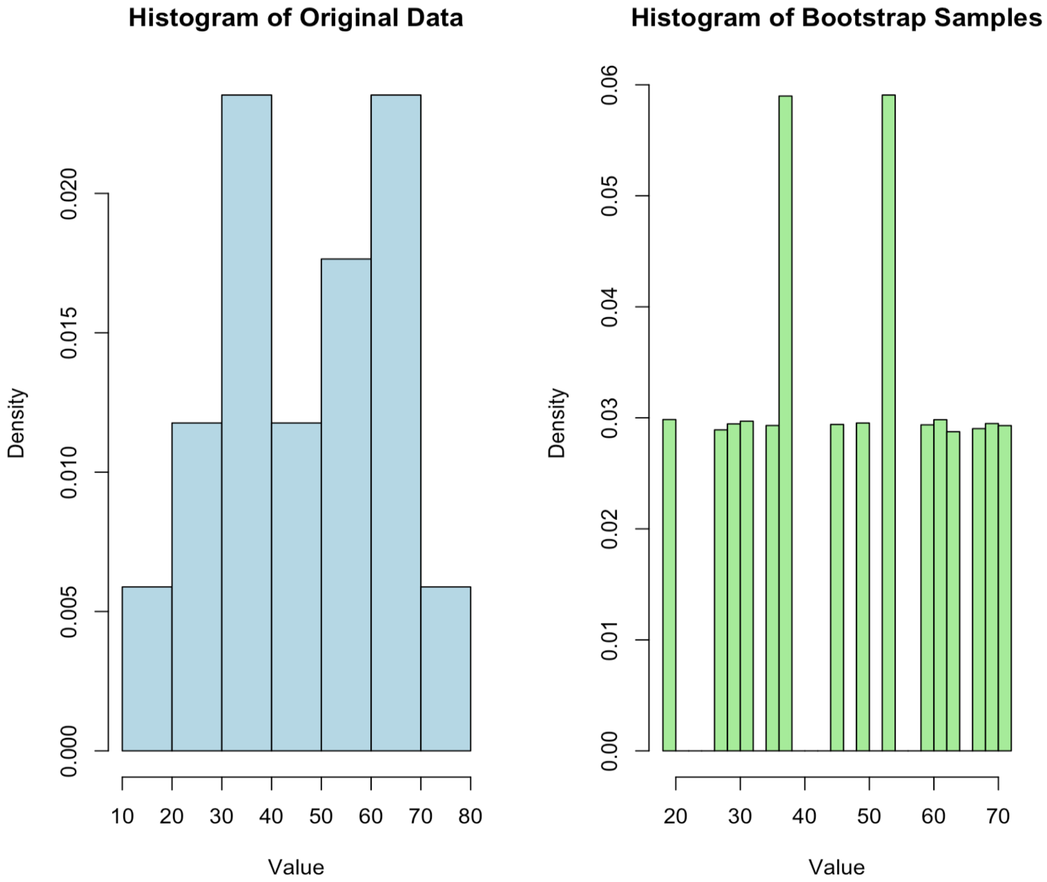

Figure 3: Histogram analysis of original and bootstrap sample data.

4.8.1 Original data

The histogram on the left represents the original dataset. It is evident from the plot that the data is somewhat uniformly spread across a range of values, with slight peaks around certain intervals. This distribution provides an initial understanding of the underlying characteristics of the data. However, given the limited sample size, it is challenging to draw definitive conclusions about the population distribution based solely on this histogram.

4.8.2 Bootstrap samples

The histogram on the right represents the distribution of means calculated from 1000 bootstrap samples. Unlike the original dataset, the bootstrap distribution exhibits a more consistent pattern with multiple distinct peaks. These peaks arise from the repetitive sampling process inherent in the bootstrap method. Each peak corresponds to the mean values generated from the repeated resampling of the original data. The bootstrap histogram demonstrates that the means are clustered around the original sample mean, with variability that is reflective of the bootstrap resampling process.

4.8.3 Interpretation

The comparison between the original data histogram and the bootstrap histogram provides important insights into the stability and variability of the sample estimates. The bootstrap histogram, with its narrower and more defined peaks, suggests that the bootstrapping technique effectively captures the variability of the sample mean and provides a more precise estimate of the underlying population mean. This visual comparison also highlights the utility of bootstrap methods in making inferences when the sample size is small, as it allows for a better understanding of the distributional properties of the estimator.

The consistency observed in the bootstrap histogram reinforces the reliability of the bootstrap estimates and suggests that the confidence intervals and hypothesis tests conducted in this analysis are well-supported by the data.

4.9 Real-world application

We apply the bootstrapping method to a real-world problem, specifically focusing on the analysis of a pre-election opinion poll from the 1972 U.S. presidential race between Nixon and McGovern. The poll surveyed 2000 participants, with the following results (see Table 1) (Wikipedia contributors, 2024):

Table 1: Election forecast based on 2000 surveys.

	Nixon	McGovern	Overall
Votes	1240	760	2000
Proportion	62%	38%	100%

The actual election results closely matched the poll's prediction, with Nixon receiving 60.67% of the votes and McGovern 37.52%.

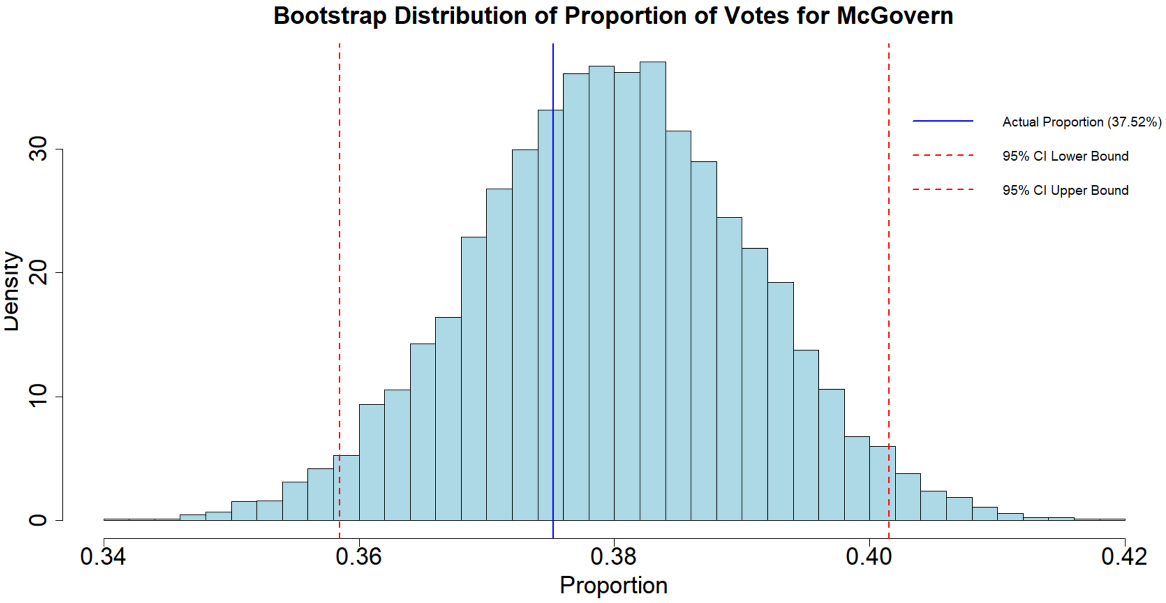

Figure 4: Bootstrap distribution of the proportion of votes for McGovern.

To assess the reliability of the poll's prediction, we used our custom-built bootstrapping method to create a distribution of the proportion of votes for McGovern. Figure 4 shows the bootstrap distribution along with the 95% confidence interval (indicated by the red dashed lines) compared to the actual election result.

The close alignment between the bootstrap confidence interval and the actual election result suggests that the poll was a reliable predictor of the election outcome. This demonstrates the effectiveness of our bootstrapping method in providing robust estimates and confidence intervals, even when the underlying distribution is not known.

4.10 Runtime comparison

In this section, we analyze the runtime performance of two different bootstrap methods: a custom implementation and a built-in method. The comparison was conducted by measuring the execution time required to generate bootstrap samples, with the number of samples ranging from 10^3 to 10^7.

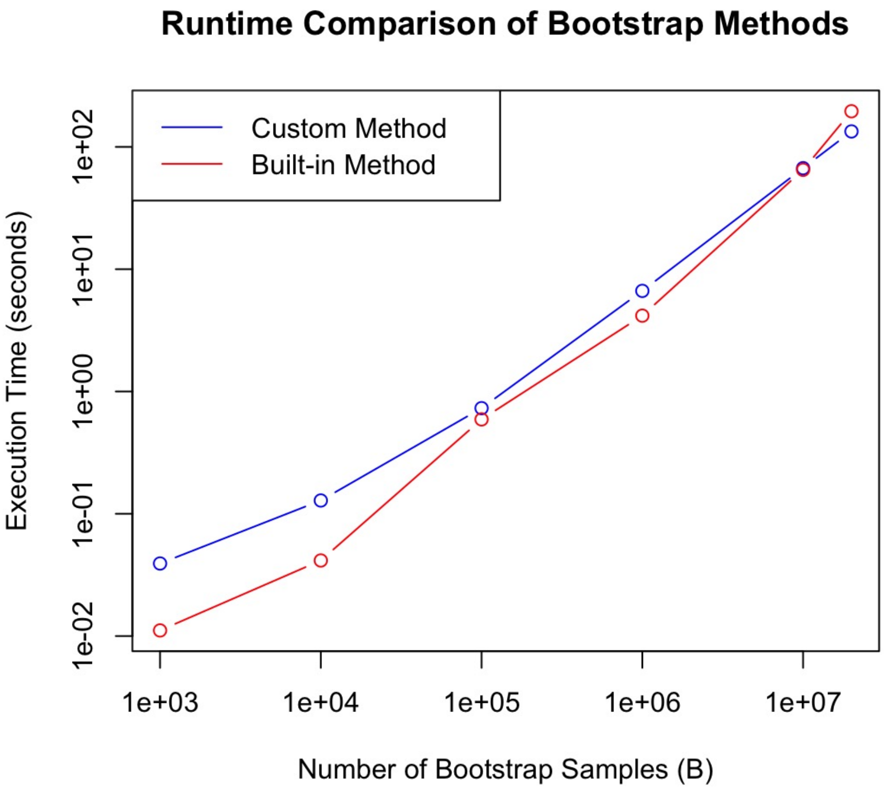

Figure 5: Runtime comparison between custom method and built-in method.

The results are illustrated in Figure 5, where the execution time (in seconds) is plotted on the y-axis, and the number of bootstrap samples (B) is plotted on the x-axis. The blue line represents the custom method, while the red line corresponds to the built-in method.

As observed in Figure 5, both methods exhibit a positive correlation between the number of bootstrap samples and execution time, indicating that the computational cost increases as the sample size grows. However, the built-in method consistently outperforms the custom method across all sample sizes. Specifically, the built-in method demonstrates a lower execution time, particularly as the number of samples increases.

For smaller sample sizes (e.g., $B = 10^3$), the performance difference between the two methods is relatively minimal. However, as the number of samples increases, the custom method's execution time grows more rapidly. At the largest sample size tested $(B = 10^7)$, the custom method takes significantly longer, with the performance gap widening noticeably.

These findings highlight the efficiency of the built-in method, especially for scenarios requiring a large number of bootstrap samples. The built-in method's superior runtime performance suggests it is a more suitable choice for computationally intensive tasks, where reducing execution time is critical. Consequently, for practical applications involving extensive bootstrap sampling, the built-in method is recommended over the custom implementation.

5 Summary

This report investigated the application of bootstrapping methods to address the challenge of making reliable statistical inferences, particularly when dealing with small sample sizes or unknown data distributions. The research focused on estimating the mean, calculating standard errors, constructing various types of confidence intervals, and performing hypothesis tests using both synthetic and real-world data.

The most significant results include the successful estimation of mean scores with corresponding confidence intervals and the demonstration that bootstrapping can produce reliable and robust statistical estimates. Specifically, the analysis showed that bias-corrected and accelerated (BCa) confidence intervals provide the most conservative and reliable estimates, accounting for bias and skewness in the data. The hypothesis testing revealed no significant difference between the observed mean and a hypothesized value, supporting the validity of the bootstrapped estimates.

While the results were promising, it is important to note that bootstrapping relies on key assumptions, such as the data being representative of the population and the independence of observations. Violations of these assumptions could lead to biased or inaccurate results. Additionally, the computational complexity of bootstrapping was analyzed, showing that while it is an effective tool, it can be computationally intensive, especially with larger sample sizes.

Future work could explore the application of bootstrapping in more complex real-world scenarios, assess the impact of different types of data distributions, and compare bootstrapping with other non-parametric methods. Further investigations might also focus on optimizing the computational efficiency of bootstrapping methods to make them more suitable for large-scale data analysis.

Bibliography

Felix Bittmann. *Bootstrapping: An Integrated Approach with Python and Stata*. De Gruyter, 2021. doi: 10.1515/9783110693348. URL https://doi.org/10.1515/9783110693348.

Angelo J. Canty and Brian D. Ripley. *boot: Bootstrap R (S-Plus) Functions*, 2021. URL https://CRAN.R-project.org/package=boot. R package version 1.3-28.

Ton J. Cleophas, Aeilko H. Zwinderman, and Toine F. Cleophas. *The Interpretation of the P-Values*, pages 103–115. Springer Netherlands, Dordrecht, 3rd edition, 2006. ISBN 978-1-4020-4650-6. doi: 10.1007/978-1-4020-4650-6_9. URL https://doi.org/10.1007/978-1-4020-4650-6_9,[Accessed:12.05.2024].

Andy Field. *Discovering Statistics Using IBM SPSS Statistics*. SAGE Publications Ltd, London, 3rd edition, 2009. ISBN 978-1-84787-907-3.

Carlos G. Forero. *Cronbach's Alpha*, pages 1357–1359. Springer Netherlands, Dordrecht, 2nd edition, 2023. ISBN 978-94-007-0753-5. doi: 10.1007/978-94-007-0753-5_622. URL https://doi.org/10.1007/978-94-007-0753-5_622,[Accessed:12.05.2024].

David S. Moore, George P. McCabe, and Bruce A. Craig. *Introduction to the Practice of Statistics*. W. H. Freeman, 6th edition, 2007. ISBN 978-0333469996.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, New York, latest edition edition, 2004. ISBN 978-0387212395.

Wikipedia contributors. Polling for united states presidential elections, 2024. URL https://en.wikipedia.org/wiki/Polling_for_United_States_presidential_elections,[Accessed:09.08.2024].

Appendix

A Pseudocode of custom bootstrap method

Generate Bootstrap Sample

```
1. Length of sample $n$.
2. Create an empty array bootstrap_sample of size $n$.
3. Randomly select $n$ indices from the range 1 to $n$.
4. Create the Bootstrap sample using the generated indices.
5. Return the bootstrap_sample.
```

Bootstrap

```
1. Length of sample $n$.
2. Create an empty array bootstrap_stats of size $B$.
3. Loop $b$ from 1 to $B$:
    a) Generate a bootstrap sample using Generate Bootstrap Sample.
    b) Compute the statistic of the bootstrap sample.
    c) Store the statistic in bootstrap_stats[b].
4. Return bootstrap_stats.
```