Parametric Predictive Bootstrap and Test Reproducibility

ABDULRAHMAN ALDAWSARI

Parametric Predictive Bootstrap and Test Reproducibility

Abdulrahman Mohammed Aldawsari

A Thesis presented for the degree of
Doctor of Philosophy

Statistics
Department of Mathematical Sciences
University of Durham
England

April 2023

Dedicated to

My son Mohammed
who makes my life convivial

My family
for encouraging and prayers

My Friends
for their believes and wishes

Parametric Predictive Bootstrap and Test Reproducibility

Abdulrahman Mohammed Aldawsari

Submitted for the degree of Doctor of Philosophy
April 2023

Abstract

Bootstrap methods have become one of the most widely used statistical techniques due to their simplicity and good properties. In this thesis, we introduce a novel bootstrap method which we call the parametric predictive bootstrap (PP-B). The PP-B method relies on parametric models, and it is primarily designed for predictive inference. In the PP-B method, a single observation is sampled from the assumed distribution with estimated parameters based on an available data set of size n. Then, this observation is added to the data and the process is repeated, now with $n+1$ observations. This process continues to sample in total m values in the same way, each observation being added to the data and re-estimating the parameters before sampling the next observation. The PP-B sample consists of m newly drawn observations and excludes the n original data observations. The performance of the PP-B method is studied on finite and infinite data ranges, and compared to other bootstrap methods via simulations, which show that it works well as a method for predictive inference. The PP-B method is applied to a range of scenarios to evaluate its performance. It relies on an assumed parametric model and we examine how it performs when the model is misspecified.

A hypothesis test is one of the most important tools in the practical application of statistics. Statistical hypothesis tests can have different results when they are repeated. The reproducibility probability of hypothesis tests has gained increasing attention due to its importance in evaluating the variability and the stability of test results. The PP-B method is presented for the reproducibility probability (RP) of some parametric tests. Test reproducibility is naturally regarded as a predictive inference problem, which is consistent with the PP-B method. The explicitly predictive nature of PP-B provides an appropriate formulation for inferring RP, as the nature of RP is explicitly predictive as well. The performance of PP-B for RP is compared with the nonparametric predictive inference bootstrap method, which also has a predictive nature but does not assume a parametric model.

Declaration

The work in this thesis is based on research carried out in the Department of Mathematical Sciences at Durham University. No part of this thesis has been submitted elsewhere for any degree or qualification, and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

First and foremost, I am genuinely grateful to my Almighty God ”Allah” for the countless blessings he has bestowed upon me generally throughout my life, and particularly in accomplishing this thesis.

I would like to offer my sincere gratitude to my supervisors Prof. Frank Coolen and Dr. Tahani Coolen-Maturi for their valuable advice and support, more particularly for the fact that they were always more than happy to help me with any problem I encountered during my PhD research.

I am extremely grateful to my family and friends for their support, love and prayers. I would like to express my gratitude to the Saudi Arabian Cultural Bureau in London and Prince Sattam Bin Abdulaziz University in Saudi Arabia for supporting me financially and granting me a scholarship to complete my studies abroad. Also, I would like to thank Durham University for providing the facilities that have enabled me to study smoothly. Finally, my thanks go out to everyone who has assisted, encouraged, or contributed to my educational progress in any way.

Notation

NPI	Nonparametric predictive inference.
$A_{(n)}$	Hill's assumption.
$\underline{P}$	NPI lower probability.
$\overline{P}$	NPI upper probability.
NPI-RP	NPI for reproducibility probability.
$\underline{RP}$	NPI lower reproducibility probability.
$\overline{RP}$	NPI upper reproducibility probability.
EB	Efron's bootstrap.
PB	Parametric bootstrap.
NPI-B	Nonparametric predictive inference bootstrap.
PP-B	Parametric predictive bootstrap.
n	Sample size.
N	The number of simulations.
CP	Coverage proportion for the confidence interval.
AW	Average width of confidence intervals.
LC	A percentile prediction interval used by Lu and Chang.
CP_{LC}	Coverage proportion for the prediction interval based on LC method.
AW_{LC}	Average width of prediction intervals based on LC method.

MT	A percentile prediction interval used by Mojirsheibani and Tibshirani.
CP_{MT}	Coverage proportion for the prediction interval based on MT method.
AW_{MT}	Average width of prediction intervals based on MT method.
PP-B-RP	Reproducibility probability using parametric predictive bootstrap.
NPI-B-RP	Reproducibility probability using nonparametric predictive inference bootstrap.
PP-BF-RP	Reproducibility probability using parametric predictive bootstrap with fixed variance.

Contents

Chapter 1

Introduction

1.1 Overview

Measuring the uncertainty of a sample estimate is an important aspect of statistical inference. Bootstrap methods are sampling techniques to quantify the uncertainty of sample estimates [10]. They have been applied to a wide range of statistical problems due to their simplicity of implementation and the possibility of providing good approximate results for sample estimates. A researcher may use the bootstrap method to avoid performing complicated mathematical derivations or, in some instances, to provide a solution where no analytical answer is possible [40]. The bootstrap method has contributed to resolving problems such as the estimation of the standard error for the statistical estimators. The standard error can be used to evaluate the accuracy of an estimator, but for the majority of statistical estimators, there are no mathematical formulas to estimate the standard error. The bootstrap exploits the power of the computer to assess the statistical accuracy of complicated procedures. Additionally, the bootstrap method is capable of determining the confidence interval for a parameter of interest efficiently. The use of the bootstrap method has been extended to many problems, including hypothesis testing, because of its simplicity to implement and its good performance.

The first presentation of the bootstrap method was in a Stanford University technical report by Bradley Efron in 1977, followed by his famous paper in the Annals of Statistics in 1979 [10, 30]. Many efforts have been made to popularise the bootstrap method in the statistical community, such as Dianconis and Efron [28], Efron [31], and Efron and Gong [35]. There are many modifications to Efron's bootstrap, such as double bootstrap, smooth bootstrap, and Bayesian bootstrap that have been presented in the literature, see e.g. [4, 21, 66]. Bootstrap methods have been introduced for different types of data, e.g. real data [44], right-censored data [2], and ordinal data [7]. Chernick [8] described bootstrap methods along with examples and applications such as hypothesis testing, confidence intervals, regression and time series. As a result, the importance of the bootstrap approach has been widely recognized. There were a few resampling techniques that predate Efron's bootstrap, such as the jackknife method which was presented by Quenouille in 1949 [32]. The method was initially developed by Quenouille for nonparametric estimation of bias. The jackknife method will be discussed in Section 2.5.

This thesis presents a new bootstrap method, the parametric predictive bootstrap, which we denote by PP-B. This method is completely based on parametric models and it is mainly designed for inferences aimed at prediction. The proposed bootstrap method will be evaluated in a range of scenarios in order to investigate its performance in estimation and predictive inference. Hypothesis testing is an important tool in practical statistics. In the application of statistical hypothesis tests, the results of the tests can differ each time they are repeated. The reproducibility probability (RP) is an important factor in the reliability of the statistical test results. It is naturally considered as a predictive problem which is well aligned to the predictive nature of PP-B. In this thesis, we study RP within a frequentist statistical framework from a prediction point of view.

1.2 Outline of thesis

This thesis is organised as follows: Chapter 2 introduces preliminary materials from the literature relevant to this thesis. Three bootstrap methods are presented first, followed by a brief introduction to the topic of reproducibility probability for tests. Finally, we discuss some methods for comparing the performance of different bootstrapping techniques.

Chapter 3 introduces PP-B and the differences between Efron's, parametric, and NPI bootstrap methods. A comparison of three bootstrap methods from the literature with PP-B is presented using the methods described in Chapter 2. PP-B will be compared to other bootstrap methods using confidence intervals and prediction intervals in terms of the coverage probability through simulations. The comparison is carried out with finite and infinite range of data.

Chapter 4 presents a summary of four parametric tests, one-sample t-test, two-sample t-test, Welch's t-test and F-test. This is followed by introducing the PP-B method for the reproducibility of these parametric tests, as well as comparing its performance with the NPI-B for test reproducibility (Bootstrap-RP). The simulation studies are performed to get an insight into how different bootstrap methods perform for RP of various tests. A comparison of Bootstrap-RP methods with NPI-RP from the literature is presented, as well as an explanation of why the Bootstrap-RP method is used rather than the NPI-RP method. We will illustrate the consistency of the proposed bootstrap method for RP with NPI-RP. Some results of this chapter were presented at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics in the US (online), October 2021 and the Royal Statistical Society Conference in the UK, September 2022.

In Chapter 5, we examine the performance of PP-B regarding the assumed parametric model. The PP-B method relies on the assumption of a parametric model, and we investigate how it performs when the PP-B samples are generated using a different model from the original data. The best bootstrap method is determined by considering a global measure of coverage accuracy as proposed by Banks [4]. This technique creates confidence regions for the bootstrap confidence interval to see the coverage probabilities for a specific parameter of interest. Then, the chi-square goodness of fit test is used to measure the discrepancy in coverage probability. Also, Banks' comparison method is applied to the bootstrap prediction interval. In addition, we investigate how the fixed bootstrap variance of PP-B impacts the reproducibility probability of one-sample t-test.

In Chapter 6, we end with conclusions and some interesting topics to extend the research presented in this thesis. Some results of Chapters 3 and 5 were presented at the International Conference of the ERCIM WG on Computational and Methodological Statistics in the UK, December 2022. The appendix includes extra simulation results. In this thesis, calculations were performed using the statistical software program R.

Chapter 2

Preliminaries

This chapter provides a review of the basic concepts from literature relevant to the topics considered in this thesis. First, we introduce the methodology of nonparametric predictive inference which is used in the topics of bootstrap and test reproducibility. Then, we introduce three bootstrap methods from the literature to compare later with our PP-B method. After that, we present a general review of reproducibility. Finally, we provide some of the most commonly used ways of comparing bootstrap performance in estimation and predictive inference.

2.1 Nonparametric predictive inference (NPI)

The nonparametric predictive inference (NPI) method has been developed during the past two decades for a wide range of applications and problems in statistics along with a variety of data types. NPI is a statistical technique based on Hill's assumption $A_{(n)}$ that makes inferences on a future observation based on past data observations [11, 12]. Hill [45, 46, 47] introduced the assumption $A_{(n)}$ for prediction of one future observation X_{n+1} with no prior knowledge about the underlying distribution. Suppose that $x_1, ..., x_n$ are the observed data corresponding to real-valued and exchangeable random quantities $X_1, ..., X_n$. Let $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ be the ordered observations and define $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$ for ease of notation. For one future observation X_{n+1}, the assumption $A_{(n)}$ is:

$$P(X_{n+1} \in I_i) = \frac{1}{n+1} \tag{2.1}$$

where $I_i = (x_{(i-1)}, x_{(i)})$ and $i = 1, \ldots, n+1$. The assumption $A_{(n)}$ states that the future observation X_{n+1} is equally likely to fall in each open interval $(x_{(i-1)}, x_{(i)})$. These intervals were created by the previous n observations between consecutive order statistics of the given sample.

The assumption $A_{(n)}$ itself is not sufficient to derive precise probabilities for any event of interest, but it can be used to derive bounds (lower and upper) of probabilities, which are called imprecise probabilities. The NPI approach is introduced by Coolen and Augustin [3, 18] which uses lower and upper probabilities for events of interest considering future observations based on Hill's assumption. The lower probability is the maximum lower bound for the precise probability for the event and denoted by $\underline{P}(\cdot)$. The upper probability is the minimum upper bound for the event and denoted by $\overline{P}(\cdot)$. The NPI lower and upper probabilities become precise probability if they are equal $\underline{P}(\cdot) = \overline{P}(\cdot)$, $0 \leq \underline{P}(\cdot) \leq \overline{P}(\cdot) \leq 1$. The NPI lower and upper probabilities for the event $X_{n+1} \in B$, where $B \subset \mathbb{R}$ are:

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1}|\{i : I_i \subseteq B\}| \tag{2.2}$$

$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1}|\{i : I_i \cap B \neq \emptyset\}| \tag{2.3}$$

The lower probability (2.2) is the total probability mass assigned to intervals I_i that are completely contained within B, and the upper probability (2.3) is taking into account all probability masses assigned to intervals that can be in B.

Sequential application of the assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$ can be used to generalize NPI for $(m \geq 1)$ future real-valued observations based on n real data observations. These assumptions imply that all $\binom{n+m}{n}$ possible different orderings of the m future observations among the n data observations are equally likely to appear, with no further assumptions made on where future observations will be within each of these intervals I_i [17]. The NPI approach is considered for statistical inference, e.g. acceptance sampling [14], precedence testing for two groups [20], and accuracy of diagnostic tests [19].

2.2 Bootstrap methods

In this section, we describe three different bootstrap methods: Efron's bootstrap (EB), parametric bootstrap (PB), and nonparametric predictive inference bootstrap (NPI-B). These bootstrap methods will be compared with the parametric predictive bootstrap (PP-B) introduced in this thesis. The classical Efron bootstrap is a nonparametric sampling technique that does not make any assumptions about how observations are distributed. In contrast, the parametric bootstrap requires assumptions regarding the distribution of the data. The nonparametric predictive inference bootstrap is formulated for predictive inference and does not use an assumed parametric model. The PP-B is similar to NPI-B in terms of focusing on prediction but it requires assumptions about data distribution.

2.2.1 Efron's bootstrap

The bootstrap method has become an essential technique for researchers because of its good properties and general applicability to a variety of statistical situations. The standard version of the bootstrap method is introduced by Efron [40], which is a resampling technique from the original data set. This bootstrap method uses the empirical distribution to quantify the uncertainty of sample estimates. The basic idea of Efron's bootstrap (EB) is resampling with replacement from the original observations repeatedly, where each observation has equal probability of being selected during the resampling process [51]. It has been widely used in applied statistics as it relies on few mathematical assumptions and can be implemented easily using statistical software. It is important to note that EB makes no assumptions regarding the distribution of observations [43, 61].

Suppose that there is a random sample $x_1, x_2, \ldots, x_n$ from an unknown distribution F, and we want to estimate the parameter of interest $\theta(F)$, e.g. the mean or variance, by the statistic T. The bootstrap method can be used to construct the sampling distribution of any statistic. A bootstrap sample is denoted by $X^* = (x_1^*, x_2^*, \ldots, x_n^*)$, which consists of members of the original data set $X = (x_1, x_2, \ldots, x_n)$. It is obtained by randomly sampling n times with replacement from the original sample. The size of a bootstrap sample can be chosen differently from the original sample size. The basic bootstrap method generates an empirical estimate of the sampling distribution of the statistic (bootstrap distribution). The procedure involves drawing a large number of samples from the observations and determining the statistic for each sample. The statistic's sampling distribution can be estimated by the relative frequency distribution of these statistics. The bootstrap distribution typically mirrors the shape of the actual sampling distribution resulting from the sampling process.

A point worth noting here is that some of the observations will be repeated once or more in a bootstrap sample, which makes them different from the original sample. Also, certain observations may not appear at all in a particular bootstrap sample. Consequently, there will be a variation of the values for the parameter of interest. We should draw large numbers of bootstrap samples to approximate the variation of a sampling distribution. The EB method is described in many references with examples and applications, e.g. Berrar [6], Davison and Hinkley [21], and Efron [34]. The idea of bootstrap has been applied to a variety of statistical inferences. For example, Rosenkranz [62] estimated the bias of treatment effect estimators using the bootstrap method.

2.2.2 Parametric bootstrap

The parametric bootstrap (PB) method assumes that the data come from a known distribution with unknown parameters. In this method, samples are drawn from the assumed distribution with the estimated parameters instead of resampling with replacement from the original data. The idea of the PB method is to estimate the parameters of the assumed distribution using available data, and then to generate a number of PB samples from the assumed distribution with the estimated parameters [43, 55]. The PB method requires knowledge of the data distribution and can contain observations that weren't included in the original sample, but this method may produce misleading results if the assumed model is wrong. Conversely, EB method does not assume a distribution for the data, all observations are included in the original sample, and tied observations occur. The PB method can be used in situations where some knowledge about the form of the underlying population is available.

2.2.3 Nonparametric predictive inference bootstrap

Coolen and Binhimd [16] introduced a predictive bootstrap method based on NPI, called nonparametric predictive inference bootstrap (NPI-B). The NPI-B method involves creating $n+1$ intervals between the n ordered observations of the original data, then selecting one of these intervals randomly. The first observation is drawn uniformly from the selected interval and then added this observation to the original data, resulting in $n+1$ observations. This leads to creating a partition consisting of $n+2$ intervals, from which the second observation is sampled. The process continues until m observations are drawn, where m is predefined. These m observations constitute one NPI-B sample (which of course does not include the n original data observations). In NPI-B, all possible orderings of the new observations among the past observations are equally likely to occur. NPI-B's sampling method, which involves drawing each observation from the intervals in the partition created by combining the n original observations together with all previously drawn observations belonging to the same bootstrap sample, leads to more variation in bootstrap samples than Efron and parametric bootstrap samples.

One observation is sampled uniformly from each chosen interval when applying NPI-B. However, it cannot be sampled uniformly from an open-ended interval, e.g., data defined on the whole real line lead to the first and last intervals in the form of $(-\infty, x_{(1)})$ and $(x_{(n)}, +\infty)$. Coolen and Binhimd [16] suggest to use the tail of a Normal distribution for real-valued data, and the tail of an Exponential distribution for non-negative real-valued data. It is important to note that the conditional tail distribution is only used to sample an observation from open-ended intervals, otherwise the observation is sampled uniformly from finite intervals. The NPI-B algorithm for real-valued data on finite and infinite intervals is as follows:

1. Create $(n+1)$ intervals between the n ordered observations $x_{(0)}, x_{(1)}, x_{(2)}, \ldots,$ $x_{(n)}, x_{(n+1)}$, where $x_{(0)}$ and $x_{(n+1)}$ are the end points of the possible data range: $(x_{(0)}, x_{(1)}), (x_{(1)}, x_{(2)}), \ldots, (x_{(n-1)}, x_{(n)}), (x_{(n)}, x_{(n+1)})$.

2. Select one of the $n+1$ intervals randomly, each with equal probability, and sample one future observation uniformly from this selected interval.

   (a) We sample the future value uniformly for any finite interval.

   (b) For the case with data on the whole real line $(-\infty, +\infty)$: If the chosen interval is $(-\infty, x_{(1)})$ or $(x_{(n)}, +\infty)$, we sample the future value from the tail of Normal distribution with mean $\mu = \frac{x_{(1)}+x_{(n)}}{2}$ and standard deviation $\sigma = \frac{x_{(n)}-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ^{-1} indicating the inverse function of a standard normal cumulative distribution function.

   (c) For the case with data on the $(0, +\infty)$: If the chosen interval is $(x_{(n)}, +\infty)$, we sample the future value from the tail of Exponential distribution with rate $\lambda = \frac{\ln(n+1)}{x_{(n)}}$.

3. Add this sampled observation x_1^* to the data; increase n to $n+1$.

4. Repeat Steps 1-3, now with $n+1$ data, to obtain a further future value. This is continued to sample m future observations from the intervals in the partition created by combining the n original observations with all previously drawn observations that belong to the bootstrap sample. These m drawn observations $(x_1^*, x_2^*, \ldots, x_m^*)$ form one NPI-B sample of size m.

5. Repeat Steps 2-5 to obtain B of NPI-B samples of size m.

2.3 Reproducibility

The term "reproducible" for the findings of a study refers to the ability of results gained via an experiment or a statistical analysis of a data set to be reproduced when the study is replicated. It is considered one of the key concepts of scientific methods and gives investigators confidence in knowing precisely what has been achieved. Over the last few years, reproducibility has obtained increasing attention and several scientific journals have launched a campaign to raise awareness on reproducibility issues, titled "Journals unite for reproducibility" [57]. Many institutional drug agencies such as the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA) usually require at least two adequate and well-controlled clinical trials for evaluating the efficacy and safety of a new drug product before marketing approval [56]. The main purpose of conducting a second clinical trial is to support the effectiveness of a certain treatment and to investigate whether the clinical result of the first trial is reproducible in the second clinical trial.

Statistical tests are the tools employed as experimental evidence to support the effectiveness of the treatment. The results of statistical hypothesis tests can be different each time the tests are repeated. The topic of reproducibility probability (RP) of a hypothesis statistical testing framework was first addressed by Goodman, who pointed out that there seemed to be a misunderstanding about the meaning of a statistical p-value [42]. According to Goodman, the replication probability can be used to show that the p-value may exaggerate the evidence against the null hypothesis. In a later extensive discussion of Goodman's paper, Senn [64] disagrees with Goodman's statement that "p-values overstate the evidence against the null hypothesis" and he emphasizes the difference between the p-value and the RP. However, Senn agreed with Goodman about the importance of reproducibility of test results. Although acknowledging a natural relation between RP and the p-value, it is necessary to consider the difference between them. The p-value is an indication of the strength of the statistical evidence, and the smaller p-value in the case of rejecting the null hypothesis, the larger one would expect the RP to be. Senn [64] additionally discussed issues with the reproducibility of tests in real world situations where a repeated test may be performed in varying circumstances or be carried out by a different team of analysts.

The RP of a test is the probability that the same test outcome, either rejection of the null hypothesis or not, would be reached if the test were repeated based on an experiment performed in the same way as the original experiment. It indicates the reliability of the result of a statistical hypothesis test. The focus is usually on the reproducibility of tests that led to the rejection of the null hypothesis, as significant effects in clinical trials typically lead to new treatments in medical applications. According to Begley and Ellis [5], researchers from California attempted to confirm published findings in preclinical cancer research from 53 'landmark' studies, but they managed to obtain the same scientific findings in only 6 cases. Also, they report similar studies conducted by a team at Bayer HealthCare in Germany were able to reproduce only about 25% of the same scientific findings. Begley and Ellis [5] concentrate on improving the preclinical environment and building a stronger system in detail without discussing the statistical techniques implemented in preclinical tests. They provide recommendations to enhance the credibility of studies, such as avoiding the publication propensity to only positive results, and they emphasize the importance of RP for more reliability of medical tests.

During recent years, there has been growing interest in the RP due to an important aspect of the practical relevance of test results. Shao and Chow [65] present three approaches to evaluate RP under several different study designs commonly used in clinical trials: the estimated power approach, the lower confidence bound of power estimate, and the Bayesian approach. They use the available test data from the previous trial(s) to estimate the power of a future test, and they consider the lower confidence bound of this power estimate as a more conservative approach for the RP, in particular when the clinical result from the first trial is highly significant. Shao and Chow [65] introduced a concept of RP for a given clinical trial and both argued that a single clinical trial is sufficient if the statistical result from the first clinical trial is evaluated to be strongly reproducible. They study the generalization of the clinical results from one patient population to a different patient population and also adjust the sample size for the second trial. De Martini [25] used the power of the test as an estimate of RP to evaluate the results for a large class of parametric tests. In addition, he proposed to define the statistical tests themselves using the estimated RP. The power approach was also followed by De Capitani and De Martini [22, 23, 24] to study the RP estimation for various nonparametric tests, such as Wilcoxon signed rank test, sign test, Kendall test and binomial test. The power of a test is defined as the probability of rejecting the null hypothesis if an alternative hypothesis is true. The estimation of RP using the power approach is somewhat restrictive because it focus only on the cases where null hypothesis is rejected which is not consistent with the natural interpretation of test reproducibility. Also, the repeated application of the test, which would lead to different data, is not taken into consideration.

Miller [58] emphasizes the importance of recognising the distinction between two scenarios for test repetition. The first scenario is a general repetition of tests by other researchers working independently, in which conditions may differ from the original experiment. The second scenario is an individual repetition of tests by the same researcher, where the tests are performed under exactly the same circumstances as the original experiment and test. Miller [58] is doubtful about the ability to make inferences that are useful and precise enough from the initial experiment, in particular when the true effect size is unknown, and consequently the power of the test is unknown. We will concentrate on the second scenario, 'individual repetition of tests' in Miller's terminology, to investigate RP in this thesis, because it is conceivable to derive meaningful frequentist inferences in this scenario. We define statistical reproducibility for a test as the probability that the same test outcome would be reached if the test were repeated in the same way as the original experiment.

A new perspective on test reproducibility was presented by Coolen and Binhimd [15], using the nonparametric predictive inference (NPI) framework of frequentist statistical methods. Coolen and Binhimd [15] introduce NPI for reproducibility probability (NPI-RP) of some nonparametric tests, namely the sign test, Wilcoxon's signed-rank test and the two-sample rank-sum test. This method considers the test result for a predicted future sample of the same size as the original sample ($m = n$) to reflect the nature of reproducibility. The NPI approach focuses explicitly on future observations and uses few modelling assumptions, which causes imprecision in this process that can be quantified by the use of lower and upper probabilities. NPI-RP considers reproducibility of tests from the perspective of prediction instead of estimation, which is the substantial difference between NPI-RP and estimated power approach of Shao and Chow [65]. Also, it presents for any possible results of the original test, including both rejection and non-rejection of the null hypothesis. The focus is usually on the reproducibility of tests that led to the rejection of the null hypothesis, as significant effects in clinical trials typically lead to new treatments in medical applications. However, we believe that the reproducibility of tests that do not produce significant effects should also be considered for a complete view. NPI for reproducibility probability has been applied to a range of nonparametric tests, such as the quantile test, and the precedence test [13].

The general idea of the NPI-RP approach considers $\binom{n+m}{n}$ different orderings of the m future real valued observations among the n data observations, where these orderings all have the same probability $\binom{n+m}{n}^{-1}$ to occur. The different orderings of the m future observations among the n data observations are denoted by O_j for $j = 1, \ldots, \binom{n+m}{n}$. The number of future observations in the interval $(x_{(i-1)}, x_{(i)})$ can be expressed by $s_1^j, \ldots, s_{n+1}^j$ according to ordering O_j, where $s_i^j \geq 0$ and $\sum_{i=1}^{n+1} s_i^j = n$. We do not know precise values of the future data for any future ordering O_j, but specify the number s_i^j of observations in the interval $(x_{(i-1)}, x_{(i)})$, for each $i = 1, \ldots, n+1$. There is no additional assumption for these future observations, so they can have any value within the specific interval. The NPI-RP considers all different possible orderings of m future observations among n data observations given the observed data from the original test. The same test is performed on the future data sets as was applied to the original data and the proportion of these that lead to the same conclusion as the original test is investigated. A more detailed explanation will be provided in Section 4.6 within the context of NPI-RP for the likelihood ratio test. The NPI lower and upper reproducibility of the test are denoted by $\underline{RP}$ and $\overline{RP}$, respectively. A limitation of the NPI-RP method is that computation becomes impractical if we consider a large sample size. For example in the case of $m = n = 30$, there are $\binom{n+m}{n} = \binom{30}{15} = 155117520$ possible different orderings of the m future observations among the n real data observations, which must be computed to derive NPI lower and upper reproducibility values. To overcome this difficulty, Coolen and Binhimd [16] use a bootstrap technique for finding the RP of tests. They introduced an NPI bootstrap method to predict future samples, and they demonstrated how this method avoids the complex calculations encountered with the NPI-RP method.

The NPI-RP approach is only feasible for small data sets to compute the exact NPI lower and upper reproducibility probabilities. To overcome computational limitations associated with large sample sizes, Coolen and Marques [17] propose a sampling methodology based on sampling future orderings. They introduced an alternative computational method for the reproducibility of likelihood ratio tests with the test criterion in terms of the sample mean. The sampling procedure for the orderings meets the requirements of simple random sampling (SRS). The probability of each ordering being selected must be the same at each selection, and independent of the other selections. A large number of orderings is sufficient to eliminate any potential differences between sampling with and without replacement and for simplicity, a sample of orderings with replacement is used. An easy way to implement sampling of orderings by random sampling of a vector of integers $(r_1, \ldots, r_n)$, with $r_1 \geq 1, r_l > r_{l-1}$ for all $l = 2, \ldots, n$ and $r_n \leq 2n$. Among the $n + m$ combined data and future observations, r_l is considered to be the rank of the jth ordered data observation. Defining $s_l^j = r_l - r_{l-1} - 1$ for $l = 1, \ldots, n+1$, with a sampled vector $(r_1, \ldots, r_n)$, where $r_0 = 0$ and $r_{n+1} = 2n + 1$, thus creating the jth sampled future ordering in the SRS process. This process ensures that each possible ordering has an equal probability of being selected and independent of the other selections, which satisfies the requirements for SRS. The NPI-RP method can be applied to a wide range of statistical tests other than likelihood ratio tests, as a result of sampling orderings for estimating NPI lower and upper RPs.

2.4 Measures of statistical accuracy

A comparison between bootstrap methods can be performed by computing some statistical accuracy measures, e.g. variance, standard error, bias, root mean squared error, and mean absolute error. It is important to explain the rationale for choosing these measures. The variance is used to measure the variability between bootstrap methods. Standard error, bias, root mean squared error, and absolute error are the most commonly used measures of statistical accuracy for estimators. Suppose that $x_1, x_2, \ldots, x_n$ represent observations corresponding to independent and identically distributed random variables $X_1, \ldots, X_n$ with distribution function F. Let θ be the parameter of interest, e.g. the mean or variance, which can be estimated by the statistic T. The bootstrap method can be used to estimate the sampling distribution of T. There are two main types of bootstrap, namely parametric and nonparametric. The parametric bootstrap is useful for comparison to nonparametric analyses when some knowledge about the underlying population is available. The nonparametric bootstrap, as in EB, is typically applied when F is unknown. The distribution F is estimated by the empirical distribution function $\hat{F}$, which puts probability $1/n$ on each of the observed values. Here we show the steps of the nonparametric bootstrap (EB) as it is the standard version of the bootstrap method. The following is the algorithm of EB [30]:

1. Construct the empirical probability distribution function $\hat{F}$ by putting probability $1/n$ to each value $x_1, x_2, \ldots, x_n$, $\hat{F}(x) = \sum_{i=1}^{n} I(x_i \leq x)/n$, where $I(x_i \leq x)$ is the indicator function which is 1 if $x_i \leq x$ and 0 otherwise.

2. Draw B independent random samples of size n by sampling with replacement from the original data set.

3. Compute the statistic of interest T for each bootstrap sample to obtain $T_1^*, T_2^*, \ldots, T_B^*$.

4. Construct the empirical distribution of $T_1^*, T_2^*, \ldots, T_B^*$ by putting probability $1/B$ at each one of them, which can be used to approximate the sampling distribution of θ.

The bootstrap estimate for the standard error $\hat{se}_B$ can be computed by the sample standard deviation of $T_1^*, T_2^*, \ldots, T_B^*$ as follows:

$$\hat{se}_B = \left[\frac{\sum_{j=1}^{B} (T_j^* - T^*(.))^2}{B-1} \right]^{1/2} \tag{2.4}$$

where, $T^*(.) = \sum_{j=1}^{B} T_j^*/B$.

The main advantage of the bootstrap method is that it can be used to estimate the standard error for any estimator. The variance is computed by the square of Equation (2.4). Another useful measure of statistical accuracy is bias, which is the difference between the expectation of an estimator T and the quantity being estimated θ,

$$bias_F = bias(T, \theta) = E_F(T) - \theta \tag{2.5}$$

An estimator with good properties is desirable such as small standard error and small bias. If $bias(T, \theta) = 0$, then T is called an unbiased estimator of θ; otherwise, it is a biased estimator of θ. The bias of estimators play an important role in statistical theory and a large bias is usually an undesirable characteristic of an estimator's performance. The bootstrap can be used to estimate the bias of any estimator T by substituting F by $\hat{F}$ in Equation (2.5), leading to bootstrap estimate of bias:

$$bias_{\hat{F}} = E_{\hat{F}}(T^*) - T^0 \tag{2.6}$$

where T^0 is the computed value of statistic T based on the original sample. The bootstrap estimate of bias can be approximated by generating independent bootstrap samples and evaluating the statistic T^* for each one, then the bootstrap expectation $E_{\hat{F}}(T^*)$ can be computed by the average $T^*(.) = \sum_{j=1}^{B} T_j^*/B$. We obtain the bootstrap estimate of bias, based on B bootstrap samples, by substituting $T^*(.)$ for $E_{\hat{F}}(T^*)$ in Equation (2.6) as follows:

$$\widehat{bias}_B = T^*(.) - T^0 \tag{2.7}$$

The bootstrap estimate of bias can be computed by applying exactly the bootstrap algorithm for estimating standard error except that we calculate $T^*(.) - T^0$ at the last step rather than $\hat{se}_B$. A measure of accuracy that uses both standard error and bias is the root mean square error (RMSE) of an estimator T for θ as

$$RMSE = \sqrt{\hat{se}_B^2 + \widehat{bias}_B^2} \tag{2.8}$$

The absolute error is defined as the difference between a measured value and an actual value. The absolute error for each bootstrap sample is computed, then the average of these values is used to calculate the mean absolute error:

$$MAE = \frac{1}{B} \sum_{j=1}^{B} |T_j^* - T^0| \tag{2.9}$$

2.5 Confidence intervals

We begin this section by providing a general review of confidence intervals, followed by a discussion of how a bootstrap technique can be used to construct different confidence intervals. A $100(1-2\alpha)\%$ confidence interval for the parameter θ is an interval constructed from a random sample, such that if we were to repeat the experiment a large number of times, the interval would contain the true value of θ in $100(1-2\alpha)\%$ of the cases. It is important to note that the interval will depend on the value of the estimate $\hat{\theta}$ and the sampling distribution of the estimator. The sample size, confidence level, and the variability in the sample are all factors that influence the width of the interval. The larger samples produce narrower confidence intervals when all other factors are equal, while a higher confidence level or greater variability in the sample produces wider confidence intervals when all other factors are equal. In one sample case, we have a random sample observations $x_1, x_2, ..., x_n$ from unknown distribution F. Assume we estimate the parameter of interest θ and denoting it by $T = \hat{\theta}$ and assigning $\widehat{se}$ as an estimate of standard error of $\hat{\theta}$. We need to determine the sampling distribution of the estimator $\hat{\theta}$. Under most circumstances the distribution of $\hat{\theta}$ becomes more and more normal when the sample size n grows larger [40]. So, when the sample size n grows large, the distribution of $\hat{\theta}$ becomes asymptotically normal with a mean θ and variance $\widehat{se}^2$, which means $\hat{\theta}\dot{\sim}N(\theta, \widehat{se}^2)$ or equivalently

$$Z = \frac{\hat{\theta}-\theta}{\widehat{se}}\dot{\sim}N(0,1) \tag{2.10}$$

Now, we find endpoints $\hat{\theta}_{lo}$ and $\hat{\theta}_{up}$ which are random values defined on a random sample such that:

$$P(\hat{\theta}_{lo} \leq \theta \leq \hat{\theta}_{up}) = 1-2\alpha$$

There are an infinite number of $100(1-2\alpha)\%$ confidence interval for the parameter of interest θ, but our aim is to choose the best confidence interval available. A narrower confidence interval is not necessarily a better one, but the length and shape are only significant if the coverage probabilities are accurate. If we take approximation (2.10) to be exact and let $z^{(\alpha)}$ is the $100.\alpha$th percentile point of standard normal distribution, we obtain

$$P(z^{(\alpha)} \leq \frac{\hat{\theta}-\theta}{\widehat{se}} \leq z^{(1-\alpha)}) = 1-2\alpha$$

By rearranging this statement around the parameter of interest θ we obtain,

$$P(\hat{\theta} - z^{(1-\alpha)} \cdot \widehat{se} \leq \theta \leq \hat{\theta} - z^{(\alpha)} \cdot \widehat{se}) = 1 - 2\alpha$$

The confidence interval is obtained as follows:

$$(\theta_{lo}, \theta_{up}) = (\hat{\theta} - z^{(1-\alpha)} \cdot \widehat{se}, \hat{\theta} - z^{(\alpha)} \cdot \widehat{se}) \tag{2.11}$$

It is called the standard confidence interval with coverage probability $1 - 2\alpha$ or confidence level $100(1 - 2\alpha)\%$. Since $z^{(\alpha)} = -z^{(1-\alpha)}$ we can be expressed (2.11) in a more familiar form

$$\hat{\theta} \pm z^{(1-\alpha)} \cdot \widehat{se} \tag{2.12}$$

The standard confidence interval from the normal distribution is valid for a large number of data as $n \to \infty$, but for small data it is an approximation. In 1908, a better approximation using Student's t-distribution was derived by Gosset [70]. The Student's t interval is obtained using $t_{n-1}^{(1-\alpha)}$ which represents $(1 - \alpha)$th percentile of Student's t-distribution with $n - 1$ degree of freedom.

$$\hat{\theta} \pm t_{n-1}^{(1-\alpha)} \cdot \widehat{se} \tag{2.13}$$

The Student's t distribution is similar to the normal distribution with its bell shape but has larger tails and the population variance is unknown. It gets more close to the normal distribution as sample size increases, and the difference between them becomes negligible, but it is important with small sample sizes. When $n \geq 20$ the percentiles of student's t_n distribution do not differ much from those of $N(0, 1)$. The Student's t-distribution depends on the normality assumption, and we expect poor performance estimate for standard interval and Student's t interval in case of a non-normal distribution.

Now we explore four different estimations of confidence intervals using the bootstrap method [40]. We will investigate each method separately and explain the circumstances for each of them to provide an accurate estimate. The first approach is the bootstrap-t confidence interval that estimates the distribution of Z directly from the data by generating B bootstrap samples and for each one we find

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{se^*_B}(b)} = \frac{T_j^* - T}{\widehat{se^*_B}(b)} \tag{2.14}$$

where, $\hat{\theta}^*(b)$ is the value of $\hat{\theta}$ based on bootstrap sample x^{*b} and $\widehat{se}^*_B(b)$ is the estimated standard error of $\hat{\theta}^*$ for the bootstrap x^{*b}. The αth percentile of $Z^*(b)$ is estimated by the value $\hat{t}^{(\alpha)}$ as follows:

$$\frac{\#\{Z^*(b) \leq \hat{t}^{(\alpha)}\}}{B} = \alpha \tag{2.15}$$

where $\hat{t}^{(\alpha)}$ is the αth percentile of $Z^*(b)$ across all bootstrap samples and $\hat{t}^{(1-\alpha)}$ is $(1-\alpha)$th percentile. For example, if $B = 1000$ and $\alpha = 0.05$, then $\hat{t}^{(\alpha)}$ is the 50th largest value of the $Z^*(b)$ and $\hat{t}^{(1-\alpha)}$ is the 950th largest value of the $Z^*(b)$. If $B \cdot \alpha$ is not an integer, we assuming $\alpha \leq 0.05$ and let $k = \lfloor (B+1)\alpha \rfloor$ is the largest integer less than or equal $(B+1)\alpha$, then we define α and $1-\alpha$ by the kth largest and $(B+1-k)$th largest value of $Z^*(b)$, respectively. Therefore, a $100(1-2\alpha)\%$ bootstrap-t confidence interval is obtained as following:

$$(\hat{\theta} - \hat{t}^{(1-\alpha)} \cdot \widehat{se}^*_B, \hat{\theta} - \hat{t}^{(\alpha)} \cdot \widehat{se}^*_B) \tag{2.16}$$

The disadvantage of the bootstrap-t interval is that it cannot be trusted to construct a confidence interval for all problems and tend to give erratic results in small samples. Another more reliable approach than bootstrap-t to estimate the confidence interval using the bootstrap technique is the percentile interval, which depend on the percentiles of the bootstrap distribution of a statistic. The $100(1-2\alpha)\%$ percentile interval is giving by:

$$\begin{aligned} \hat{\theta}^{*(\alpha)}_B &< \theta < \hat{\theta}^{*(1-\alpha)}_B \\ T^{*(\alpha)}_B &< \theta < T^{*(1-\alpha)}_B \end{aligned} \tag{2.17}$$

where $T^{*(\alpha)}_B$ is the $100 \times \alpha$th percentile of the T^*_j values, that means the $B \times \alpha$th of the ordered list of the B replications of T^* and it is likewise for $T^{*(1-\alpha)}_B$ indicate the $100 \times (1-\alpha)$th percentile of the T^*_j values. For example, if $B = 1000$ and $\alpha = 0.05$, then lower endpoint of the percentile interval $T^{*(\alpha)}_B$ is the 50th ordered value of replications and upper endpoint of the percentile interval $T^{*(1-\alpha)}_B$ is the 950th ordered value of replications. If $B \times \alpha$ is not an integer, we follow the same procedure of bootstrap-t interval as discussed earlier.

The percentile interval of θ is well aligned with a standard normal interval created by transforming θ in a manner that normalises the distribution and then mapped to the θ scale [40]. The percentile method incorporates the correct transformation automatically without requiring the statistician to know the transformation. Also, the percentile interval has transformation respecting (invariant) property, that means any (monotone) parameter transformation $\phi = m(\theta)$ of percentile interval is simply the percentile interval for θ mapped by $m(\theta)$ [40]:

$$[\hat{\phi}^{(\alpha)}, \hat{\phi}^{(1-\alpha)}] = [m(\hat{\theta}^{(\alpha)}), m(\hat{\theta}^{(1-\alpha)})]$$

The percentile method gives good performance if $\hat{\theta}$ is an unbiased estimator, and the transformation has an approximately normal distribution with a constant variance [36].

The last two methods to construct a confidence interval are an improved version of the percentile method known as BC and BCa. Efron introduced the bias-corrected (BC) bootstrap interval which is an improved version of the percentile interval [31]. The BC method designed to work well when there exist a monotone transformation $\phi = m(\theta)$ and the estimator $\hat{\phi} = m(\hat{\theta})$ is approximately normal distribution with the mean $\phi - z_0\sigma$ and constant standard deviation such that

$$\hat{\phi} \sim N(\phi - z_0\sigma, \sigma^2) \tag{2.18}$$

where, the parameter z_0 is the bias correction that can be obtained by bootstrap method and σ is the standard deviation of $\hat{\phi}$ that does not depend on ϕ, which signifies that the standard deviation is a fixed constant.

The basic idea of the BC method is to build a confidence interval for a monotone transformation ϕ and then, transform these back to θ scale by using the inverse of the monotone transformation. In other words, a confidence interval for ϕ is constructed with the help of Formula (2.18) by using the same logic that gave standard normal interval as follows

$$(\hat{\phi} + z_0\sigma) \pm z^{(1-\alpha)}\sigma$$

Then, the confidence interval is converted back for θ by using the inverse transformation $\theta = m^{-1}(\phi)$. The advantage of the BC approach is that all of this is accomplished automatically through bootstrap computations, without needing explicit knowledge of the transformation m [33]. The value of the bias-correction z_0

counts the possible bias in T as an estimate of θ. It can be obtained uses the ratio of bootstrap replications less than the original estimate T,

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{T_j^* < T\}}{B}\right) \tag{2.19}$$

where, $\Phi^{-1}(\cdot)$ indicates to the inverse function of a standard normal cumulative distribution function, e.g. $\Phi^{-1}(0.95) = 1.645$. The value of $\hat{z}_0$ equal zero if exactly half of the bootstrap replications T_j^* values are less than or equal to the original estimate T. The BC interval of intend coverage $1 - 2\alpha$ is obtained by

$$\text{BC} : (\hat{\theta}_{lo}, \hat{\theta}_{up}) = (T_B^{*(\alpha_1)}, T_B^{*(\alpha_2)}) \tag{2.20}$$

where,

$$\alpha_1 = \Phi\left(2\hat{z}_0 + z^{(\alpha)}\right) \tag{2.21}$$

$$\alpha_2 = \Phi\left(2\hat{z}_0 + z^{(1-\alpha)}\right) \tag{2.22}$$

Here α_1 is the lower endpoint percentile while α_2 is the upper endpoint percentile. $z^{(\alpha)}$ is the 100αth percentile point of a standard normal distribution, and $\Phi(\cdot)$ is the standard normal cumulative distribution function. The BC method helps in correcting deficiencies of the percentile method when the value of $\hat{z}_0$ is non-zero by changing the percentiles used for the BC endpoints. However, if the value $\hat{z}_0$ equal zero, then Formulas (2.21) and (2.22) will be

$$\alpha_1 = \Phi(z^{(\alpha)}) = \alpha \quad \text{and} \quad \alpha_2 = \Phi(z^{(1-\alpha)}) = 1 - \alpha \tag{2.23}$$

Hence, the BC interval in Equation (2.20) will be the same as the percentile interval in Equation (2.17), meaning both methods will give the same estimate of the confidence interval when there is no bias.

The last method we discuss for constructing a confidence interval is the bias-corrected and accelerated (BCa) bootstrap interval. This method is more difficult to compute than the BC, but yields more accurate confidence intervals with less restrictive assumptions where it does not assume $\hat{\phi} = m(\hat{\theta})$ has a constant standard deviation such that

$$\hat{\phi} \sim N(\phi - z_0\sigma_\phi, \sigma_\phi^2) \tag{2.24}$$

where, σ_ϕ is the standard deviation of $\hat{\phi}$ that does depend on ϕ as follows: $\sigma_\phi = 1+a\phi$ which is a linear function of ϕ. The acceleration a is a small constant which explains how the standard deviation of $\hat{\phi}$ varies with ϕ.

The BCa method shares similar concepts with the BC method except that it does not assume $\phi = m(\theta)$ having a constant standard deviation. So, it builds a confidence interval for ϕ and then converts it back to an interval for θ by the inverse transformation $\theta = m^{-1}(\phi)$ which is done automatically without the need to know the transformation [37, 39]. Estimating the bias-correction from bootstrap distribution is elaborated previously in BC method. There are several methods for calculating acceleration, we will use the easiest way and computed separately from the bootstrap distribution using the jackknife approach [29]. The jackknife method is a resampling technique that pre-dates the bootstrap method, and both have similarities [40]. The jackknife method has a specific number of resamples which depends on the sample size, while in bootstrap there is no agreed number of replications, and it depends on the interest estimation of the confidence interval or standard error and so on. The jackknife is usually less computationally demanding than the bootstrap. The resample procedure of the jackknife method is based upon the sequential removal of one observation from the dataset n times. The size of the jackknife sample is $n-1$. Assuming we have a sample $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and we estimate the interested parameter $\hat{\theta} = t(\mathbf{x})$, the jackknife sample omitting one observation at a time is

$$\mathbf{x}_{(i)} = (x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \tag{2.25}$$

for $i = 1, 2, \dots, n$, called jackknife samples and $T_{(i)}$ is ith jackknife replication of T. We can estimate the acceleration using the jackknife method in the following formula [40]:

$$\hat{a} = \frac{\sum_{i=1}^{n}(T_{(\cdot)} - T_{(i)})^3}{6\{\sum_{i=1}^{n}(T_{(\cdot)} - T_{(i)})^2\}^{3/2}}; \qquad T_{(\cdot)} = \sum_{i=1}^{n} T_{(i)}/n \tag{2.26}$$

The BCa interval of intend coverage $1-2\alpha$ is obtained by

$$\text{BCa} : (\hat{\theta}_{lo}, \hat{\theta}_{up}) = (T_B^{*(\alpha_1)}, T_B^{*(\alpha_2)}) \tag{2.27}$$

where,

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right) \tag{2.28}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right) \tag{2.29}$$

However, if the value of $\hat{a}$ is equal to zero, then Formulas (2.28) and (2.29) will be

$$\alpha_1 = \Phi\left(z_0 + \tfrac{z_0 + z^{(\alpha)}}{1 - 0(z_0 + z^{(\alpha)})}\right) = \Phi\left(2z_0 + z^{(\alpha)}\right) \tag{2.30}$$

$$\alpha_2 = \Phi\left(z_0 + \tfrac{z_0 + z^{(1-\alpha)}}{1 - 0(z_0 + z^{(1-\alpha)})}\right) = \Phi\left(2z_0 + z^{(1-\alpha)}\right) \tag{2.31}$$

Therefore, the BCa interval in Equation (2.27) will be the same as the BC interval in Equation (2.20), implying that both methods will give the same estimate of the confidence interval.

2.6 Prediction intervals

We begin this section by providing an overview of prediction intervals and explaining how to construct prediction intervals using the bootstrap technique. Lu and Chang [54] used the bootstrap method to construct a prediction interval for one or more future values from a Birnbaum-Saunders distribution. They constructed the prediction interval using the bootstrap percentile method with bootstrap calibration. The Birnbaum-Saunders distribution is used in reliability applications to model failure times. They assumed that a random sample $x_1, \ldots, x_n$, is taken from Birnbaum-Saunders distribution function F with parameters α and β. The density of F is defined by

$$f(x;\alpha,\beta) = \frac{1}{2\alpha\beta\sqrt{2\pi}}\left[\left(\frac{x}{\beta}\right)^{-\frac{1}{2}} + \left(\frac{x}{\beta}\right)^{-\frac{3}{2}}\right] exp\left[-\frac{1}{2\alpha^2}\left(\frac{x}{\beta} - 2 + \frac{\beta}{x}\right)\right],$$

where $x > 0$ and the two parameters of Birnbaum–Saunders distribution $\alpha, \beta > 0$.

A bootstrap sample of size n, $x_1^*, \ldots, x_n^*$ is n random values drawn with replacement from $x_1, \ldots, x_n$, each with a probability of $1/n$, to construct the estimated distribution F^*. In this case, the bootstrap sample is considered as a sample of the unknown distribution. Then, generate $y_1^*, \ldots, y_m^*$ from the estimated distribution F^*, where m is the number of future observations. Thereafter, the mean of $y_1^*, \ldots, y_m^*$ is obtained and denoted by $\bar{y}_m^*$. Repeat the previous procedure B times to obtain B values of $\bar{y}_m^*$, denoted by $\bar{y}_m^*(1), \ldots, \bar{y}_m^*(B)$. Then, construct $100(1-2\alpha)\%$ prediction interval for the mean of future observations $\bar{x}_m$ as:

$$\left(\bar{y}_{m,B}^{(\alpha)}, \bar{y}_{m,B}^{(1-\alpha)}\right) \tag{2.32}$$

where the lower endpoint $\bar{y}_{m,B}^{(\alpha)}$ is the $B \times \alpha$th value in the ordered list of the B replications of $\bar{y}_m^*$ and the upper endpoint $\bar{y}_{m,B}^{(1-\alpha)}$ is the $B \times (1-\alpha)$th value in this ordered list. If $B \times \alpha$ is not an integer, the same procedure of bootstrap-t interval is used as discussed in previous section (use the largest integer).

Lu and Chang [54] investigate the performance of the bootstrap prediction intervals for a single future observation and for the mean of five future observations through simulations. They draw a sample of size $n+m$ from the Birnbaum-Saunders distribution $x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m}$, where $x_1, \ldots, x_n$ represents the past sample and $x_{n+1}, \ldots, x_{n+m}$ represents the future sample. Then, find the observed mean of m future observations $x_{n+1}, \ldots, x_{n+m}$, $\bar{x}_m$. The prediction interval for the mean of future observations is constructed by drawing the bootstrap sample $x_1^*, \ldots, x_n^*$, then generating from them $y_1^*, \ldots, y_m^*$ and finding $\bar{y}_m^*$. Repeat this $B = 1000$ times to have the list of B values $\bar{y}_m^*(1), \ldots, \bar{y}_m^*(B)$ in order to construct the prediction interval for $\bar{x}_m$ as described earlier. In the case of prediction for a single future observation, we draw the bootstrap sample $x_1^*, \ldots, x_n^*$, then generate y_1^* from them and repeat this $B = 1000$ times to have the list of B values $y_1^*(1), \ldots, y_1^*(B)$ and construct the prediction interval for x_{n+1}. The 90% and 95% prediction intervals for a single future observation x_{n+1} and the mean of m future observations $\bar{x}_m$ are computed in the Lu and Chang study [54]. They conducted a Monte-Carlo simulation to determine the coverage probability by counting how many intervals contain x_{n+1} and $\bar{x}_m$. A percentile prediction interval used by Lu and Chang [54] is defined as the LC method.

Mojirsheibani and Tibshirani [60] introduced different ways to construct prediction intervals such as bootstrap-t, percentile and BCa prediction intervals. The percentile prediction interval is used in this thesis because the BCa prediction interval cannot be constructed for a single future observation and the bootstrap-t prediction interval is not transformation respecting. They assumed that a random sample $X = (x_1, \ldots, x_n)$ represent past sample and $Y = (y_1, \ldots, y_m)$ represent a future sample, where X and Y are independently and identically distributed with a common distribution F and $\hat{\theta} = T$ is a scalar parameter. Let F_n and F_m are the CDF's of $\hat{\theta}_n = T_n$ and $\hat{\theta}_m = T_m$, which are the estimators of a scalar parameter from the past sample and future sample respectively. Let $\hat{F}_n$ and $\hat{F}_m$ are the CDF's of $\hat{\theta}^*_n = T^*_n$ and $\hat{\theta}^*_m = T^*_m$, the bootstrap version of $\hat{\theta}_n = T_n$ and $\hat{\theta}_m = T_m$. The bootstrap samples X^* and Y^* are drawn with replacement from the past sample X. The $100(1-2\alpha)\%$ percentile prediction interval for $\hat{\theta}_m = T_m$ is:

$$(\hat{\theta}_{lo}, \hat{\theta}_{up}) = \left(\hat{F}_m^{-1}\left[\Phi(z^{(\alpha)}(1+m/n)^{1/2})\right], \hat{F}_m^{-1}\left[\Phi(z^{(1-\alpha)}(1+m/n)^{1/2})\right]\right) \tag{2.33}$$

where $\hat{F}_m$ is the bootstrap distribution of $\hat{\theta}^*_m = T^*_m$, and $z^{(\alpha)} = \Phi^{-1}(\alpha)$.

Mojirsheibani [59] studies the effects of bootstrap iteration (calibration) as a method to improve the coverage accuracy. They generated X^* and Y^* from the past sample X and then resample Y^{**} from X^*. All previous studies of constructing prediction intervals were based on Efron's bootstrap method. We refer to the percentile prediction interval recommended by Mojirsheibani and Tibshirani [60] as the MT method. In this thesis, we focus on the percentile prediction interval using MT and LC methods, but without iterated bootstrap.

This chapter provides background information for the topic discussed in this thesis by presenting the main concepts from the literature. In Chapter 3, we introduce a novel bootstrap method called parametric predictive bootstrap (PP-B). This method is presented for the RP of some parametric tests in Chapter 4. The PP-B method is primarily designed for predictive inference and it relies on an assumed parametric model. In Chapter 5, we examine how it performs when the model is incorrectly specified.

Chapter 3

Parametric Predictive Bootstrap

3.1 Introduction

This chapter introduces the parametric predictive bootstrap (PP-B), a novel bootstrap method for predictive inference. Additionally, we examine its performance in a range of scenarios that have been used with other bootstrap methods, as discussed in Chapter 2. The strength of estimation and prediction inference of PP-B is evaluated using measures of statistical accuracy, confidence intervals, and prediction intervals. PP-B is compared with other methods of bootstrap, described in Section 2.2, to demonstrate the differences between them.

This chapter is organized as follows: Section 3.2 introduces the idea of parametric predictive bootstrap and clarifies the differences compared to three other bootstrap methods: Efron's, parametric, and NPI-Bootstrap. Two different scenarios are considered to compare the PP-B method to other bootstrap methods using measures of statistical accuracy, confidence intervals, and prediction intervals. The first scenario generates data from a distribution with finite support. For the second scenario, data sets are generated from a distribution with infinite support. We provide a comparison of the two scenarios in Section 3.3 and Section 3.4, respectively. In Section 3.5, the performance of PP-B for estimation is compared with different types of bootstrap methods using percentile confidence intervals. In Section 3.6, we extend the comparison of the percentile prediction intervals to prediction of the future sample statistics. In Section 3.7, we present some concluding remarks of this chapter.

3.2 Parametric predictive bootstrap

In this section, we present the main idea of PP-B for real valued data, followed by a brief initial comparison with other bootstrap methods described in Section 2.2. In the PP-B method, a single observation is sampled from an assumed distribution with estimated parameters based on an original data set of size n. Then, this observation is added to the data and the process is repeated, now with $n+1$ observations. We re-estimate the distribution parameters with the new observation added to the data in order to sample the second observation. This process continues to sample m further values in the same way, each observation adding to the data and re-estimating the parameters before sampling the next one. The PP-B sample consists of these m sampled observations, so it excludes the n original data observations. The PP-B algorithm for one-dimensional real-valued data is as follows:

1. We have a random sample consisting of n observations $x_1, x_2, \ldots, x_n$ from a known distribution $F(x;\theta)$, with parameter θ.

2. The parameter θ of the assumed distribution is estimated by $\hat{\theta}$ from the available data, using maximum likelihood estimation (MLE) or any other estimation method.

3. Sample one future observation x_1^* randomly from the fitted distribution $F(x;\hat{\theta})$.

4. Add x_1^* to the data giving data set $(x_1, x_2, \ldots, x_n, x_1^*)$; increase n to $n+1$.

5. Repeat Steps 2-4, now with $n+1$ data, to obtain a further future value. This is continued to sample m observations in total, with each one added to the data and the parameter re-estimated before sampling the next observation. These sampled observations $x_1^*, x_2^*, \ldots, x_m^*$ are a PP-B sample of size m.

6. Repeat Steps 2-6 to obtain B of PP-B samples of size m.

As a consequence of the method of sampling observations in PP-B, with sampled observation added to the data set and the parameter estimated before sampling the next one, the bootstrap samples show more variation than the EB and PB samples. The method for sampling observations in NPI-B, with each observation drawn from the intervals created by combining the n original observations with all previously drawn observations belonging to the same bootstrap sample, also causes more variation in the bootstrap samples than the EB and PB samples. In the EB and PB methods, all observations are sampled based on the original data only. EB depends on resampling with replacement from the original data, where each value of the original data set has the same probability of being chosen by random selection during the resampling process [38]. The PB method assumes the data to come from a known distribution with unknown parameters. The parameters of the assumed distribution are estimated from the available data, then observations are sampled from the assumed distribution with the estimated parameters in order to obtain PB sample [43]. The bootstrap samples in PP-B, NPI-B, and PB are not restricted to already observed values, whereas all observations in EB samples are in the original sample.

We give a brief initial comparison of variations in bootstrap samples for each bootstrap method and leave a more detailed comparison for the following sections. We compute the variance for a statistic of interest T using the bootstrap technique to measure the spread of these statistic values based on the bootstrap samples. The bootstrap variance for the mean and variance is estimated using $B = 1000$ bootstrap samples for each bootstrap method. We generate an original sample of size n from N(0,1) and then apply different bootstrap methods $B = 1000$ times. The mean and variance of each bootstrap sample are computed, and then we estimate the variance based on different bootstrap methods. We repeat this procedure with different original sample sizes $n = 5, 25, 100, 200, 500$ from N(0,1), as well as with Exp(0.5). The same data sets of each sample size from N(0,1) and Exp(0.5) are used with all bootstrap methods. It is important to note that the bootstrap samples for each method have the same size as the original sample.

Table 3.1 shows the estimate of variance using different bootstrap methods for the mean and variance. The results were approximated to four decimal digits, but we used additional digits with some values to make the results more informative

(a) N(0,1)

method	statistics	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PP-B	mean	0.3320	0.0606	0.0146	0.0083	0.0041
	variance	0.5491	0.1201	0.0249	0.0151	0.0084
NPI-B	mean	0.6067	0.0892	0.0165	0.0085	0.0040
	variance	2.9020	0.3624	0.0444	0.0172	0.0112
PB	mean	0.1981	0.0316	0.0075	0.0042	0.0020
	variance	0.0009	0.0001	0.00002	0.00001	0.000004
EB	mean	0.1515	0.0359	0.0074	0.0043	0.0021
	variance	0.1743	0.0735	0.0133	0.0071	0.0047

(b) Exp(0.5)

method	statistics	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PP-B	mean	0.4517	0.1764	0.0789	0.0386	0.0149
	variance	4.3163	2.0995	1.9951	0.8993	0.3314
NPI-B	mean	0.6725	0.2146	0.0839	0.0321	0.0140
	variance	11.7162	9.6144	4.5507	1.1390	0.5424
PB	mean	0.2385	0.0887	0.0396	0.0194	0.0075
	variance	2.6386	1.5059	1.3511	0.5997	0.2195
EB	mean	0.1211	0.0717	0.0352	0.0143	0.0070
	variance	0.1063	0.3657	0.6540	0.2132	0.1408

Table 3.1: *The bootstrap estimate of variance for the mean and variance when the original sample was from N(0,1) and Exp(0.5).*

and to avoid the inclusion of zeros ”0.0000”. PP-B and NPI-B have the largest estimated variance values for the mean and variance among these bootstrap methods, as expected due to the method of sampling observations in both methods. The results for N(0,1) show that the NPI-B method has the largest variance in all cases except for the mean when $n = 500$, in which case the PP-B method has a larger variance. Also, the NPI-B method provides the largest variance in most cases of the mean and all cases of the variance for Exp(0.5), followed by the PP-B method. The NPI-B method has a larger variance in most cases compared to the PP-B method due to the assumption of a parametric model in the PP-B method.

3.3 Finite support scenario

In this section, we assess the performance of the parametric predictive bootstrap method (PP-B) regarding estimation and predictive inference and we compare it with other bootstrap methods. The comparison is based on measures of statistical accuracy and confidence intervals in order to investigate the performance of different bootstrap methods as an estimation approach. Additionally, prediction intervals are used to examine how the bootstrap methods perform in predictive inference. For the first scenario, we use distributions with finite support such as the uniform and Beta distributions. When NPI-B is applied to a finite data range, we sample uniformly across all intervals as discussed in Section 2.2.3.

3.3.1 Measures of statistical accuracy

We gave an overview of some measures of statistical accuracy for estimators in Section 2.4. The standard error, bias, root mean squared error, and mean absolute error are computed to evaluate the statistical accuracy of estimators using the bootstrap method. These measures of statistical accuracy are used to assess the performance of different bootstrap methods as an estimation approach. We investigate the PP-B's performance through simulation studies and compare it with the performance of EB, PB, and NPI-B. The smaller values of statistical accuracy measures are considered good characteristics of an estimator. We use the uniform distribution in the first scenario as an example of a distribution with finite support. The probability density function of the continuous uniform distribution is as follows

$$f(x) = \begin{cases} \dfrac{1}{b-a} & ; \quad x \in [a,b] \\ 0 & ; \quad x < a \text{ or } x > b \end{cases} \tag{3.1}$$

The data range of uniform distribution will be determined based on the parameters a and b.

Statistical accuracy measures are computed for the mean and variance using $B = 1000$ bootstrap samples for each bootstrap technique. Efron and Tibshirani [40] claim that the estimation of standard error tends not to need more than 200 replications, but confidence interval construction requires 1000 replications. An original sample of size n is generated from Uniform (2,3), and then applied different bootstrap techniques $B = 1000$ times. We compute the mean and variance for each bootstrap sample. Then, measures of statistical accuracy are estimated for the mean and variance based on different bootstrap methods, as discussed in Section 2.4. We repeat this procedure with different original sample sizes $n = 5, 25, 100, 200, 500$. For each sample size, we considered the same data sets from Uniform(2,3) with all bootstrap methods. The estimation results of statistical accuracy measures using different bootstrap methods for the mean and variance are shown in Tables 3.2 and 3.3, respectively. Note, the bootstrap samples for each method have the same size as the original sample. Also, the parameters of the uniform distribution are estimated using the method of moment estimation (MME) when applying PP-B and PB.

We begin by comparing the estimation results of statistical accuracy measures for the mean based on different bootstrap methods. PP-B has the largest value of the standard error when $n = 25, 100$, otherwise NPI-B provides a larger standard error value compared to the other bootstrap methods. PB gives the smallest value of the standard error in all cases, except for $n = 500$, where EB has the smallest value. The absolute value of bias for NPI-B is larger than for PP-B in all cases except when $n = 5, 500$, where PP-B has a larger value. PP-B has the smallest absolute value of bias when $n = 100, 200$, otherwise PB and EB have a smaller absolute bias value. PB provides a smaller absolute bias value than EB in all cases except when $n = 5, 500$. The RMSE and MAE of PP-B are smaller than those of NPI-B when $n = 5, 200, 500$. However, the RMSE and MAE values associated with PB and EB are smaller compared to the other bootstrap methods.

We also compare the estimation results of statistical accuracy measures for the variance based on different bootstrap methods. PB and EB generally produce

method	measures	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PP-B	$\hat{se}_B$	0.1524	0.0820	0.0370	0.0259	0.0177
	bias	-0.0042	-0.0016	-0.00029	-0.000001	-0.00031
	RMSE	0.1525	0.0820	0.0370	0.0259	0.0177
	MAE	0.1209	0.0650	0.0298	0.0207	0.0142
NPI-B	$\hat{se}_B$	0.1702	0.0799	0.0369	0.0267	0.0181
	bias	-0.0025	-0.0038	-0.0004	-0.0003	0.0002
	RMSE	0.1702	0.0800	0.0369	0.0267	0.0181
	MAE	0.1390	0.0642	0.0296	0.0215	0.0145
PB	$\hat{se}_B$	0.1176	0.0588	0.0264	0.0185	0.0125
	bias	-0.0026	-0.0008	-0.00034	-0.0001	-0.00028
	RMSE	0.1176	0.0588	0.0264	0.0185	0.0125
	MAE	0.0946	0.0472	0.02131	0.0148	0.0100
EB	$\hat{se}_B$	0.1181	0.0589	0.0266	0.0191	0.0122
	bias	-0.00004	0.0014	0.0013	0.0004	0.00004
	RMSE	0.1181	0.0589	0.0266	0.0191	0.0122
	MAE	0.0974	0.0474	0.02126	0.0154	0.0099

Table 3.2: *The statistical accuracy measures for the bootstrap sample mean when the original sample was from Uniform(2,3).*

smaller values of statistical accuracy than PP-B and NPI-B. This is because PB and EB are sampled individually instead of being added to the sample data set, resulting in lower variation than PP-B and NPI-B samples. Statistical accuracy measures with smaller values are desirable for estimators. NPI-B gives the largest value of the standard error in all cases, except for $n = 25$, where PP-B has the largest value. The PP-B method has the smallest value of the standard error when $n = 5$, otherwise PB and EB have a smaller standard error value compared to the other bootstrap methods. The absolute value of bias for PP-B is the largest for all

method	measures	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PP-B	$\hat{se}_B$	0.03252	0.0203	0.0087	0.0064	0.0046
	bias	-0.0254	-0.0058	-0.00083	-0.0006	-0.00033
	RMSE	0.0412	0.0211	0.0087	0.0064	0.0046
	MAE	0.0352	0.0171	0.0070	0.0051	0.0037
NPI-B	$\hat{se}_B$	0.0538	0.0202	0.0094	0.0071	0.0048
	bias	-0.0026	-0.0007	0.00081	0.0005	0.00013
	RMSE	0.0539	0.0202	0.0094	0.0071	0.0048
	MAE	0.0445	0.0161	0.0075	0.0057	0.0038
PB	$\hat{se}_B$	0.03253	0.0155	0.0063	0.0045	0.00328
	bias	-0.0144	-0.0029	-0.0003	-0.0002	-0.00018
	RMSE	0.0355	0.0158	0.0063	0.0045	0.0033
	MAE	0.0291	0.0126	0.0050	0.0036	0.0026
EB	$\hat{se}_B$	0.0376	0.0142	0.0065	0.0047	0.00336
	bias	-0.0161	-0.0042	-0.0007	-0.0003	-0.00027
	RMSE	0.0409	0.0148	0.0065	0.0047	0.0034
	MAE	0.0324	0.0119	0.0051	0.0038	0.0027

Table 3.3: *The statistical accuracy measures for the bootstrap sample variance when the original sample was from Uniform(2,3).*

cases compared to the other bootstrap methods. PB has smaller absolute bias values in all cases except when $n = 25, 500$, where NPI-B provides a smaller value. The values of RMSE and MAE in NPI-B are largest for all cases except when $n = 25$, where PP-B has the largest value. However, PB and EB produce smaller RMSE and MAE values than other bootstrap methods.

We repeated this experiment with PP-B and PB, but the parameters of uniform distribution are estimated using maximum likelihood estimation (MLE). Table 3.4 shows the estimation results of statistical accuracy measures for the mean and vari-

(a) mean

method	measures	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PP-B	$\hat{se}_B$	0.0939	0.0548	0.0281	0.0195	0.0127
	bias	0.0888	-0.0057	-0.0156	-0.0149	0.0030
	RMSE	0.1292	0.0551	0.0321	0.0245	0.0130
	MAE	0.1064	0.0441	0.0255	0.0198	0.0105
PB	$\hat{se}_B$	0.0939	0.0548	0.0281	0.0195	0.0127
	bias	0.0888	-0.0057	-0.0156	-0.0149	0.0030
	RMSE	0.1292	0.0551	0.0321	0.0245	0.0130
	MAE	0.1064	0.0441	0.0255	0.0198	0.0105

(b) variance

method	measures	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PP-B	$\hat{se}_B$	0.0208	0.0135	0.0071	0.0050	0.0033
	bias	-0.0386	-0.0141	0.0086	0.0077	0.0021
	RMSE	0.0438	0.0195	0.0112	0.0092	0.0040
	MAE	0.0393	0.0163	0.0093	0.0080	0.0032
PB	$\hat{se}_B$	0.0208	0.0135	0.0071	0.0050	0.0033
	bias	-0.0386	-0.0141	0.0086	0.0077	0.0021
	RMSE	0.0438	0.0195	0.0112	0.0092	0.0040
	MAE	0.0393	0.0163	0.0093	0.0080	0.0032

Table 3.4: *The statistical accuracy measures for the bootstrap sample mean and variance when the original sample was from Uniform(2,3).*

ance using PP-B and PB based on MLE. It is apparent from the results that PP-B and PB have exactly the same results in all cases of the mean and variance with different measures of statistical accuracy. A uniform distribution is characterized by two parameters which can be estimated from the minimum and maximum values of data based on the MLE method. In the PP-B procedure, we add the sampled observation to the data and re-estimate the parameters in order to sample the next observation. The parameters of uniform distribution will not be changed in PP-B, as a result of using the MLE method to re-estimate the parameters after each sampled observation. This occurs due to each sampled observation being restricted between the minimum and maximum values of data and adding this drawn observation to the data set does not change the parameters of uniform distribution. Therefore, the PP-B and PB methods have the same bootstrap samples due to they have the same parameters and we use the same seeds to generate different bootstrap methods. However, the MME does not make much difference for the PB method as it provides parameter estimation close to the MLE method, in particular for large sample sizes.

3.3.2 BCa confidence interval

We employ confidence intervals to evaluate different bootstrap methods for estimation. A detailed discussion of bootstrap confidence intervals was given in Section 2.5. There are several methods for constructing confidence intervals, each different in individual assumptions and the level of difficulty of computation. They become more complicated in calculations when the assumptions are lessened. The BCa interval in Equation (2.27) is chosen to compare different bootstrap methods because it has a higher order of accuracy and transformation respecting [40]. A Beta distribution with parameters α and β is used in the first scenario as an example of a distribution with finite support. The probability density function of the Beta distribution is as follows

$$f(x) = \frac{1}{\beta(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1} \quad ; \quad x \in [0,1] \tag{3.2}$$

The simulation study is conducted to find the coverage proportion and average width of confidence intervals for the mean, variance and median. In this study, we use Beta(8,2) with different original sample sizes $n = 50, 100, 200, 400$ at confidence levels 90% and 95%. We generate an original sample of size n from Beta(8,2) and then apply different bootstrap methods $B = 1000$ times. It is important to note that the bootstrap samples for each method are the same size as the original samples. The statistics are computed for each bootstrap sample to construct BCa intervals using Equation (2.27). Then, we discover which BCa confidence intervals include the true statistics of the Beta(8,2) distribution. This procedure is repeated $N = 1000$ times in order to find the coverage proportions of different bootstrap methods. The performance assessment of each bootstrap method is based on two criteria: coverage proportion and the average width of the intervals. It is desirable to have a proportion of coverage that is close to these advertised confidence levels with a smaller average width of intervals.

Table 3.5 presents the coverage proportions and average interval widths for the mean, variance, and median based on the four bootstrap procedures. The notation CP and AW refer to the coverage proportion and average interval widths, respectively. The NPI-B method produces the largest average width of confidence intervals in all cases for these three statistics, followed by the PP-B method. As a result, over-coverage occurs in all cases of the NPI-B method, as well as for the mean and variance in the PP-B method. The sampling methods in PP-B and NPI-B, which add a sampled observation to the data set before sampling the next one, leads to more variation in the bootstrap samples, as discussed in Section 3.2. The greater variability in the sample produces wider intervals, so as a result PP-B and NPI-B lead to wider confidence intervals than other bootstrap methods. The NPI-B method produces a wider average width of intervals than the PP-B method when the sample size and confidence level of both methods are equal. We conclude that the NPI-B method has more variation than the PP-B method, as we had expected, due to the assumption of a parametric model in the PP-B method. The NPI-B method does not use an assumed parametric model, leading to greater variability compared to the PP-B method.

The method that has a coverage proportions closer to nominal coverage probability is the preferred one. PB and EB have coverage that is closer to the presumed coverage probabilities with narrower intervals on average than NPI-B and PP-B for the mean and variance. For the variance, the PB method achieved the best coverage with a nominal coverage probability of 0.95 when $n = 400$, and 90% when $n = 50$. EB gives a result of 5.1% under-coverage below the nominal level of 90% for the variance when $n = 50$. Regarding the median, the results for the PB method were far worse for under-coverage than the other bootstrap methods. The EB method pro-

(a) mean

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9890	0.9930	0.9940	0.9900	0.9720	0.9760	0.9810	0.9680
	AW	0.0946	0.0669	0.0473	0.0334	0.0790	0.0561	0.0397	0.0280
NPI-B	CP	0.9950	0.9980	0.9940	0.9930	0.9820	0.9850	0.9820	0.9760
	AW	0.1046	0.0710	0.0488	0.0339	0.0865	0.0592	0.0409	0.0285
PB	CP	0.9390	0.9380	0.9550	0.9430	0.8810	0.9010	0.9060	0.9050
	AW	0.0668	0.0474	0.0334	0.0237	0.0561	0.0398	0.0281	0.0199
EB	CP	0.9340	0.9460	0.9580	0.9420	0.8800	0.8960	0.9060	0.9030
	AW	0.0660	0.0471	0.0332	0.0236	0.0555	0.0395	0.0280	0.0198

(b) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9900	0.9870	0.9910	0.9920	0.9670	0.9630	0.9800	0.9800
	AW	0.0227	0.0147	0.0097	0.0066	0.0184	0.0120	0.0081	0.0055
NPI-B	CP	0.9990	0.9980	0.9980	0.9950	0.9970	0.9940	0.9910	0.9860
	AW	0.0303	0.0183	0.0114	0.0074	0.0225	0.0140	0.0091	0.0060
PB	CP	0.9480	0.9410	0.9520	0.9500	0.9000	0.8990	0.9040	0.9020
	AW	0.0140	0.0096	0.0066	0.0046	0.0116	0.0080	0.0055	0.0038
EB	CP	0.9170	0.9270	0.9420	0.9430	0.8490	0.8720	0.8920	0.8880
	AW	0.0129	0.0093	0.0065	0.0046	0.0108	0.0078	0.0054	0.0038

(c) median

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9260	0.9210	0.9320	0.9200	0.8760	0.8680	0.8740	0.8560
	AW	0.1075	0.0768	0.0541	0.0383	0.0909	0.0649	0.0458	0.0324
NPI-B	CP	0.9910	0.9890	0.9880	0.9930	0.9750	0.9740	0.9740	0.9790
	AW	0.1211	0.0858	0.0608	0.0427	0.1017	0.0718	0.0511	0.0359
PB	CP	0.8350	0.8330	0.8500	0.8220	0.7590	0.7590	0.7700	0.7420
	AW	0.0806	0.0579	0.0411	0.0290	0.0687	0.0493	0.0349	0.0246
EB	CP	0.9400	0.9380	0.9420	0.9370	0.8940	0.8810	0.8980	0.8850

vides the best results for the median because its coverage is closer to the presumed coverage probabilities, followed by the PP-B method. PP-B has under-coverage results in all cases of the median, even though on average it has wide intervals. In Section 3.5, we will discuss in detail why both the PP-B and PB methods show under-coverage in all cases of the median. The PP-B method does not perform well in confidence intervals, as it is not developed for estimating population characteristics, but for predictive inference. It is explicitly aimed at predictive inference, with variability in different bootstrap samples reflecting uncertainty in prediction in line with the NPI-B method.

3.3.3 LC and MT prediction intervals

In this section, a comparison of PP-B with other bootstrap methods is carried out using prediction intervals in order to investigate their performance in prediction inference. A brief overview of prediction intervals based on the bootstrap technique is presented in Section 2.6. The percentile prediction intervals are constructed based on the LC and MT methods. Here we will draw the past and future samples separately as done by Mojirsheibani and Tibshirani [60], where both samples are independent and identically distributed. A coverage proportion of the percentile prediction interval for the mean of m future observations is studied using the following processes:

1. Draw an original sample of size n from specific distribution to be the past sample, giving $X = (x_1, \ldots, x_n)$. Then, draw an original sample of size m from the same distribution to be the future sample, giving $Y = (y_1, \ldots, y_m)$. The samples X and Y are independent and identically distributed.

2. Compute the observed mean of the m future observations $\bar{y}_m = \sum_{i=1}^{m} y_i/m$.

3. Draw B bootstrap samples of size m from past sample. Then, calculate the mean for each bootstrap sample $\bar{y}_m^*$ to obtain a list of $\bar{y}_m^*(j)$ for $j = 1, \ldots, B$.

4. Construct an $100(1-2\alpha)\%$ prediction interval for the mean of m future observations $\bar{y}_m$ based on the LC and MT methods:

(a) Lu and Chang (LC) method: lower endpoint is the $\alpha \times B$th value in the ordered list of $\bar{y}^*_m(j)$ and the upper endpoint is the $(1-\alpha) \times B$th value in this list (use the largest integer if these values are not integer).

(b) Mojirsheibani and Tibshirani (MT) method:
Lower endpoint $(\hat{\theta}_{lo})$: $\hat{F}^{-1}_m \left[\Phi(z^{(\alpha)}(1+m/n)^{1/2})\right] = \hat{F}^{-1}_m [\alpha_1]$ is the $\alpha_1 \times B$th value in the ordered list of $\bar{y}^*_m(j)$.
Upper endpoint $(\hat{\theta}_{up})$: $\hat{F}^{-1}_m \left[\Phi(z^{(1-\alpha)}(1+m/n)^{1/2})\right] = \hat{F}^{-1}_m [\alpha_2]$ is the $\alpha_2 \times B$th value in the ordered list of $\bar{y}^*_m(j)$.
If $\alpha_1 \times B$ or $\alpha_2 \times B$ are not integer, use the largest integer.

5. Determine if this interval contains the mean $\bar{y}_m$ of m future observations in Step 2 and compute the width of the prediction interval for both methods.

6. Steps 1-5 are repeated N times to find the coverage proportion (number of times out of N that interval captures its corresponding future sample mean) and the average interval widths.

In the case of a single future observation ($m = 1$), the percentile prediction interval is constructed as discussed in Section 2.6.

We conduct a simulation study as shown in the steps above to investigate the coverage performance and average width of intervals for each bootstrap method. The number of simulations is set equal to $N = 1000$ and the bootstrap methods are applied to each past sample $B = 1000$ times. The percentile prediction intervals are constructed for the mean of $m = n$ future observations with various original sample sizes $n = 50, 100, 200, 400$ from Beta(3,1) at confidence levels 90% and 95%. Table 3.6 shows the coverage proportions and average width of intervals for LC and MT methods using different bootstrap methods. The notation CP_{LC} and AW_{LC} refer to the coverage proportion and average interval widths for the LC prediction interval, respectively. In the MT prediction interval, CP_{MT} and AW_{MT} represent coverage proportion and average interval widths, respectively.

First, we compare the performance of different bootstrap methods with the LC prediction interval. In all future sample sizes and confidence levels, the PP-B and

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP_{LC}	0.9510	0.9390	0.9600	0.9480	0.8950	0.8910	0.9080	0.9040
	CP_{MT}	0.9900	0.9890	0.9940	0.9940	0.9710	0.9740	0.9830	0.9750
	AW_{LC}	0.1493	0.1063	0.0755	0.0534	0.1252	0.0892	0.0634	0.0449
	AW_{MT}	0.2091	0.1484	0.1054	0.0742	0.1770	0.1256	0.0894	0.0632
NPI-B	CP_{LC}	0.9600	0.9470	0.9660	0.9450	0.9220	0.8990	0.9180	0.9020
	CP_{MT}	0.9930	0.9920	0.9940	0.9950	0.9800	0.9810	0.9840	0.9800
	AW_{LC}	0.1574	0.1090	0.0763	0.0537	0.1317	0.0915	0.0641	0.0451
	AW_{MT}	0.2211	0.1524	0.1071	0.0750	0.1868	0.1292	0.0905	0.0635
PB	CP_{LC}	0.8390	0.8220	0.8360	0.8300	0.7760	0.7440	0.7450	0.7510
	CP_{MT}	0.9460	0.9380	0.9540	0.9420	0.9030	0.8870	0.9120	0.9050
	AW_{LC}	0.1066	0.0753	0.0535	0.0378	0.0896	0.0634	0.0450	0.0318
	AW_{MT}	0.1478	0.1048	0.0744	0.0525	0.1258	0.0890	0.0632	0.0447
EB	CP_{LC}	0.8380	0.8150	0.8380	0.8380	0.7630	0.7410	0.7420	0.7450
	CP_{MT}	0.9490	0.9350	0.9510	0.9420	0.8950	0.8850	0.9110	0.9030
	AW_{LC}	0.1054	0.0749	0.0533	0.0377	0.0885	0.0630	0.0449	0.0317
	AW_{MT}	0.1461	0.1041	0.0742	0.0525	0.1245	0.0884	0.0631	0.0446

Table 3.6: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for the mean of* m *future observations from Beta(3,1), when* $m = n$.

NPI-B methods provide coverage that is close to the coverage probability. Conversely, the coverage of PB and EB is considerably below the nominal coverage probability for all cases irrespective of sample size and confidence level. They provide coverage proportions that are at least 11% lower than their nominal coverage probabilities. Bootstrap methods with a predictive nature, such as PB-B and NPI-B, perform well and provide good coverage for LC prediction intervals. PP-B has the advantage of achieving good coverage with a narrower interval, where the average interval widths for PP-B are smaller than those for NPI-B in all cases.

We also compare different bootstrap methods in terms of their performance based on MT prediction intervals. The PP-B and NPI-B methods have over-coverage in all

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP_{LC}	0.9330	0.9360	0.9450	0.9370	0.8740	0.8840	0.8860	0.8880
	CP_{MT}	0.9360	0.9380	0.9450	0.9380	0.8760	0.8850	0.8870	0.8880
	AW_{LC}	0.6943	0.6950	0.6972	0.6979	0.6112	0.6117	0.6140	0.6135
	AW_{MT}	0.6990	0.6973	0.6984	0.6985	0.6159	0.6140	0.6152	0.6141
NPI-B	CP_{LC}	0.9370	0.9450	0.9450	0.9460	0.8840	0.8880	0.8890	0.8870
	CP_{MT}	0.9420	0.9460	0.9470	0.9460	0.8910	0.8890	0.8920	0.8870
	AW_{LC}	0.7271	0.7095	0.7031	0.7005	0.6320	0.6220	0.6188	0.6164
	AW_{MT}	0.7333	0.7120	0.7043	0.7011	0.6369	0.6243	0.6200	0.6170
PB	CP_{LC}	0.9330	0.9360	0.9450	0.9370	0.8740	0.8840	0.8860	0.8880
	CP_{MT}	0.9360	0.9380	0.9450	0.9380	0.8760	0.8850	0.8870	0.8880
	AW_{LC}	0.6943	0.6950	0.6972	0.6979	0.6112	0.6117	0.6140	0.6135
	AW_{MT}	0.6990	0.6973	0.6984	0.6985	0.6159	0.6140	0.6152	0.6141
EB	CP_{LC}	0.9150	0.9330	0.9420	0.9450	0.8660	0.8760	0.8870	0.8820
	CP_{MT}	0.9170	0.9340	0.9420	0.9450	0.8700	0.8770	0.8900	0.8840
	AW_{LC}	0.6757	0.6902	0.6946	0.6957	0.6056	0.6083	0.6118	0.6120
	AW_{MT}	0.6801	0.6925	0.6958	0.6963	0.6095	0.6107	0.6131	0.6127

Table 3.7: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for a single future observation from Beta(3,1), when* $m = 1$.

cases as a result of the large average interval width in both methods. The PB and EB methods provide coverage that is closer to the presumed coverage probabilities. Although the MT method improves coverage for PB and EB, it is still possible to obtain coverage closer to the coverage probability using PP-B and NPI-B with the LC method. For example, the coverage of PP-B and NPI-B with the LC method is closer to the nominal coverage probabilities when $m = 100$ than PB and EB with the MT method.

The coverage proportion of the percentile prediction interval is studied for a single future observation using the same past sample sizes and confidence levels. The LC and MT methods achieve good coverage in all cases of different bootstrap methods as shown in Table 3.7. This occurs because the future sample size is only one, which greatly impacts the interval width. A larger sample produces narrower intervals, resulting in the single future observation having a greater average width of intervals compared to the mean of m future observations in Table 3.6. The difference of coverage between the LC and MT methods is negligible. The MT interval in Equation (2.33) includes the term $(1 + m/n)^{1/2}$, which is approximately equal to 1 if $m = 1$ and n is very large. Hence, the MT method is almost the same as the LC method, e.g. the lower and upper endpoints of the MT method when $\alpha = 0.05$ are as follows:

$$\begin{aligned}\hat{\theta}_{lo} &= \hat{F}_m^{-1}\left[\Phi(z^{(0.05)}\,(1+m/n)^{1/2}\right] \approx \hat{F}_m^{-1}\left[\Phi(z^{(0.05)})\right] = 0.05\\ \hat{\theta}_{up} &= \hat{F}_m^{-1}\left[\Phi(z^{(0.95)}\,(1+m/n)^{1/2}\right] \approx \hat{F}_m^{-1}\left[\Phi(z^{(0.95)})\right] = 0.95\end{aligned}$$

where $z^{(\alpha)} = \Phi^{-1}(\alpha)$, e.g. $z^{(0.95)} = \Phi^{-1}(0.95) = 1.645$. PP-B and PB have exactly the same result in all cases of a single future observation with the LC and MT methods. It happens due to the same seeds being used to generate different bootstrap methods, which consist of only one observation in each bootstrap sample. Also, we considered the same data sets for each sample size from Beta(3,1) with all bootstrap methods. Consequently, both PP-B and PB have the same bootstrap samples, resulting in the same coverage and interval average width.

3.4 Infinite support scenario

A comparison of PP-B's performance for estimation and prediction is presented in this section. Statistical accuracy measures and confidence intervals are used to determine how the bootstrap methods perform in estimation. We use prediction intervals to examine the performance of different bootstrap methods in predictive inference. The second scenario involves distributions with infinite support such as the Normal and Gamma distributions. With finite support, for the NPI-B method one observation is sampled uniformly from an interval and this is possible for each interval. In the case of infinite support, we will have one or more infinite intervals, making it impossible to draw one observation uniformly from such intervals. To overcome this obstacle, Coolen and Binhimd [16] propose to use the tail of a Normal distribution for general real-valued data, and the tail of an Exponential distribution for non-negative real-valued data as discussed in Section 2.2.3. These assumptions are also applied in the NPI-B method in this section.

3.4.1 Measures of statistical accuracy

The performance of different bootstrap methods with infinite support is evaluated using statistical accuracy measures. Simulation studies are carried out to study PP-B's performance, and compare it with the performance of EB, PB, and NPI-B. A normal distribution with parameters μ and σ^2 is used as an example of data from a distribution with infinite support. The probability density function of the Normal distribution is as follows

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad ; \quad x \in (-\infty, \infty) \tag{3.3}$$

The bootstrap method is used for evaluating the statistical accuracy of estimators by calculating the standard error, bias, root mean square error and mean absolute error. We generate a sample of size n from N(0,1), and apply different bootstrap techniques $B = 1000$ times. The mean and variance of each bootstrap sample are calculated, then we estimate measures of statistical accuracy based on different bootstrap methods. This procedure is repeated with different original sample sizes $n = 5, 25, 100, 200, 500$, where all bootstrap methods are applied to the same data sets from N(0,1) for each sample size n. In Tables 3.8 and 3.9, we present the estimated values of statistical accuracy measures for the mean and variance based on different bootstrap methods.

We first compare the estimation results of statistical accuracy measures for the mean based on different bootstrap methods. NPI-B gives the largest value of the standard error in all cases, except for $n = 500$, where PP-B has the largest value. PB provides the smallest value of standard error when $n = 25, 200, 500$, otherwise EB has a smaller standard error value compared to other bootstrapping methods. The absolute value of bias for PP-B is smaller than for NPI-B in all cases except when $n = 5, 200$, where NPI-B has a smaller value. The smallest absolute bias value is

method	measures	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PPB	$\hat{se}_B$	0.5762	0.2461	0.1209	0.0910	0.0637
	bias	-0.0062	0.0014	-0.0024	-0.0017	-0.0003
	RMSE	0.5762	0.2461	0.1209	0.0910	0.0637
	MAE	0.4510	0.1952	0.0971	0.0733	0.0511
NPI-B	$\hat{se}_B$	0.7789	0.2987	0.1285	0.0923	0.0632
	bias	0.0060	-0.0209	-0.0041	0.0013	-0.0005
	RMSE	0.7789	0.2994	0.1286	0.0923	0.0632
	MAE	0.5846	0.2360	0.1037	0.0734	0.0506
PB	$\hat{se}_B$	0.4451	0.1777	0.0869	0.0647	0.0452
	bias	-0.0031	0.0009	-0.0020	-0.0006	-0.0005
	RMSE	0.4451	0.1777	0.0869	0.0647	0.0452
	MAE	0.3584	0.1417	0.0699	0.0520	0.0363
EB	$\hat{se}_B$	0.3893	0.1894	0.0861	0.0658	0.0463
	bias	0.0176	0.0004	0.0047	0.0004	0.0002
	RMSE	0.3897	0.1894	0.0862	0.0658	0.0463
	MAE	0.3137	0.1503	0.0685	0.0526	0.0368

Table 3.8: *The statistical accuracy measures for the bootstrap sample mean when the original sample was from N(0,1).*

obtained by PB when $n = 5, 100$, while it is obtained by EB when $n = 25, 200, 500$. The RMSE and MAE of PP-B are smaller than those of NPI-B in most cases. Nevertheless, the RMSE and MAE values obtained with PB and EB are smaller than for the other bootstrap methods. PB has the smallest RMSE and MAE values when $n = 25, 200, 500$, otherwise EB gives smaller values for the RMSE and MAE.

Also, we compare the estimation results of statistical accuracy measures for the variance using different bootstrap methods. In general, the measures of statistical accuracy using the PB and EB have smaller values compared to PP-B and NPI-B.

method	measures	$n = 5$	$n = 25$	$n = 100$	$n = 200$	$n = 500$
PPB	$\hat{se}_B$	0.7410	0.3466	0.1577	0.1230	0.0918
	bias	-0.1138	-0.0291	0.0004	0.0031	-0.0011
	RMSE	0.7497	0.3478	0.1577	0.1230	0.0918
	MAE	0.5522	0.2725	0.1245	0.0974	0.0739
NPI-B	$\hat{se}_B$	1.7035	0.6020	0.2107	0.1310	0.1058
	bias	0.6073	0.2243	0.0728	0.0259	0.0282
	RMSE	1.8085	0.6424	0.2229	0.1335	0.1095
	MAE	1.0372	0.4524	0.1674	0.1058	0.0843
PB	$\hat{se}_B$	0.6685	0.2623	0.1112	0.0869	0.0649
	bias	0.0457	0.0049	0.0063	0.0042	0.00004
	RMSE	0.6701	0.2623	0.1114	0.0870	0.0649
	MAE	0.5146	0.2084	0.0882	0.0689	0.0520
EB	$\hat{se}_B$	0.4174	0.2710	0.1152	0.0843	0.0689
	bias	-0.1895	-0.0436	-0.0054	-0.0050	-0.0006
	RMSE	0.4584	0.2745	0.1153	0.0844	0.0689
	MAE	0.3741	0.2232	0.0923	0.0675	0.0548

Table 3.9: *The statistical accuracy measures for the bootstrap sample variance when the original sample was from N(0,1).*

The reason for this is that PB and EB sample all observations based on the original data only, leading to lower variation than PP-B and NPI-B samples. The smaller values of statistical accuracy measures are regarded as positive characteristics of estimators. The NPI-B method gives the largest value of the standard error in all cases, followed by PP-B. The smallest standard error value is obtained using PB when $n = 25, 100, 500$, otherwise EB has the smallest value of standard error. The absolute value of bias for PP-B is smaller than for NPI-B in all cases, also it has the smallest absolute bias value when $n = 100, 200$. The PB has a smaller absolute value of bias than EB in all cases except when $n = 100$, where EB has a smaller value. The PP-B method has smaller values of RMSE and MAE than the NPI-B method in all cases. However, the RMSE and MAE values obtained with PB and EB are smaller than those obtained with other bootstrap methods.

3.4.2 BCa confidence interval

The BCa interval is used to evaluate the performance of different bootstrap methods in the case of infinite support. We conduct a simulation study to find the coverage proportion and average width of intervals for three statistics: mean, variance, and median. This study uses N(3,4) with a different original sample size $n = 50, 100, 200, 400$ and confidence levels 95% and 90%. We generate $N = 1000$ data sets from a Normal distribution with mean 3 and variance 4 with a specific sample size n. Then, different bootstrap methods are applied to each data set $B = 1000$ times and compute the statistics of the bootstrap samples. Following this, we construct 1000 BCa intervals using Equation (2.27). Finally, we identify the confidence intervals that include the true statistics of N(3,4) in order to determine the coverage proportions of different bootstrap methods. The coverage proportions and average width of intervals for several statistics using different bootstrap methods are outlined in Table 3.10.

PP-B and NPI-B intervals lead to wider confidence intervals than the other bootstrap methods due to the greater variability in their bootstrap samples. Consequently, the NPI-B method has over-coverage in all cases of the three statistics, as well as for the mean and variance in the PP-B method. The method that has a coverage proportion closer to nominal coverage probability with a smaller average width of intervals is preferred. PB and EB are capable of providing good coverage with smaller average interval widths for the mean and variance than NPI-B and PP-B. In the case of the mean, the EB method has a nominal coverage probability 0.95 when $n = 200$. The PB method has the best coverage for most cases of variance. Regarding the median, the EB method provides the best results because its coverage is closer to the nominal coverage probabilities, followed by the PP-B method. The over-coverage tendency associated with PP-B is disappearing in all cases of the

(a) mean

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9970	0.9970	0.9960	0.9940	0.9800	0.9840	0.9800	0.9780
	AW	1.5540	1.1028	0.7802	0.5512	1.3004	0.9244	0.6549	0.4634
NPI-B	CP	0.9970	0.9980	0.9970	0.9950	0.9860	0.9870	0.9820	0.9840
	AW	1.6698	1.1508	0.7998	0.5596	1.3932	0.9639	0.6714	0.4700
PB	CP	0.9520	0.9520	0.9480	0.9460	0.9040	0.9070	0.8990	0.8900
	AW	1.1060	0.7810	0.5523	0.3900	0.9296	0.6568	0.4643	0.3282
EB	CP	0.9480	0.9520	0.9500	0.9440	0.9050	0.9060	0.9050	0.8940
	AW	1.0949	0.7782	0.5512	0.3898	0.9211	0.6554	0.4634	0.3277

(b) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9890	0.9900	0.9910	0.9930	0.9670	0.9750	0.9760	0.9840
	AW	5.3188	3.4592	2.3272	1.6015	4.3417	2.8529	1.9382	1.3407
NPI-B	CP	0.9910	0.9940	0.9940	0.9960	0.9770	0.9790	0.9840	0.9750
	AW	5.1264	3.4615	2.3723	1.6330	4.0912	2.8167	1.9559	1.3581
PB	CP	0.9420	0.9550	0.9520	0.9480	0.8970	0.9200	0.8940	0.8880
	AW	3.3875	2.3066	1.5983	1.1157	2.8162	1.9270	1.3398	0.9364
EB	CP	0.9160	0.9450	0.9440	0.9430	0.8640	0.9070	0.8780	0.8850
	AW	3.2190	2.2604	1.5830	1.1102	2.6974	1.8928	1.3291	0.9319

(c) median

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9170	0.9110	0.9240	0.9150	0.8580	0.8590	0.8430	0.8450
	AW	1.7279	1.2230	0.8627	0.6109	1.4576	1.0367	0.7309	0.5168
NPI-B	CP	0.9920	0.9930	0.9940	0.9950	0.9790	0.9770	0.9750	0.9790
	AW	1.9741	1.3860	0.9810	0.6925	1.6492	1.1610	0.8254	0.5819
PB	CP	0.8320	0.8230	0.8090	0.8140	0.7580	0.7310	0.7270	0.7440
	AW	1.2991	0.9182	0.6526	0.4642	1.1061	0.7835	0.5559	0.3946
EB	CP	0.9460	0.9420	0.9430	0.9480	0.9080	0.9000	0.8930	0.8840

median, even though on average it has wide intervals. The PB method has worse results for under-coverage than the other bootstrap methods for the median.

3.4.3 LC and MT prediction intervals

In this section, we evaluate the prediction performance of different bootstrap techniques using prediction intervals based on the LC and MT methods in case of infinite support. We consider the Gamma distribution with two parameters α and β as an example of a distribution with infinite support. The probability density function of the Gamma distribution is as follows

$$f(x) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} \quad ; \quad x \in (0, \infty) \tag{3.4}$$

The simulation study is performed to investigate the proportion of coverage and the average interval widths for each bootstrap method. A past sample of size n and a future sample of size m are generated independently from Gamma(6,3), and the mean of the m future observations is then determined. For each bootstrap method, we draw $B = 1000$ bootstrap samples of size m from the past sample and compute the mean of each bootstrap sample. Following this, we use Equations (2.32) and (2.33) to compute the percentile prediction interval for the mean of m future observations based on the LC and MT methods. We then determine which prediction intervals include the observed mean of the m future observations. This procedure is performed $N = 1000$ times in order to see the coverage proportion and the average interval widths. In this study, we conduct simulations with sample sizes $n = 50, 100, 200, 400$ at confidence levels 95% and 90%. In Table 3.11, we present the coverage proportions and average interval widths using the LC and MT methods based on the different bootstrap methods.

We first compare the performance of different bootstrap methods with the LC prediction interval in terms of coverage proportion and average width of intervals. PP-B and NPI-B have good coverage results for the LC method, as their coverage proportions are close to the nominal coverage probabilities. PP-B has the advantage of providing a smaller average interval width than NPI-B for all future sample sizes and confidence levels. In contrast, the results of PB and EB are worse in undercov-

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP_{LC}	0.9270	0.9330	0.9330	0.9420	0.8770	0.8790	0.8810	0.8890
	CP_{MT}	0.9830	0.9860	0.9870	0.9920	0.9630	0.9660	0.9720	0.9740
	AW_{LC}	0.6289	0.4469	0.3172	0.2243	0.5245	0.3748	0.2663	0.1884
	AW_{MT}	0.8947	0.6314	0.4438	0.3135	0.7506	0.5321	0.3757	0.2657
NPI-B	CP_{LC}	0.9580	0.9440	0.9380	0.9510	0.9050	0.9050	0.8990	0.8950
	CP_{MT}	0.9930	0.9940	0.9890	0.9930	0.9820	0.9730	0.9760	0.9790
	AW_{LC}	0.7416	0.4927	0.3344	0.2312	0.6087	0.4089	0.2797	0.1940
	AW_{MT}	1.1587	0.7270	0.4786	0.3257	0.9137	0.5933	0.3994	0.2743
PB	CP_{LC}	0.8030	0.8160	0.8240	0.8280	0.7270	0.7460	0.7520	0.7460
	CP_{MT}	0.9240	0.9250	0.9310	0.9390	0.8840	0.8820	0.8790	0.8830
	AW_{LC}	0.4473	0.3176	0.2243	0.1589	0.3759	0.2668	0.1886	0.1337
	AW_{MT}	0.6219	0.4422	0.3118	0.2211	0.5288	0.3754	0.2653	0.1877
EB	CP_{LC}	0.8100	0.8100	0.8300	0.8290	0.7270	0.7450	0.7470	0.7550
	CP_{MT}	0.9180	0.9180	0.9290	0.9380	0.8730	0.8720	0.8870	0.8860
	AW_{LC}	0.4431	0.3150	0.2237	0.1587	0.3724	0.2648	0.1882	0.1334
	AW_{MT}	0.6152	0.4387	0.3116	0.2210	0.5225	0.3726	0.2647	0.1875

Table 3.11: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for the mean of* m *future observations from Gamma(6,3), when* $m = n$.

erage for all cases of the LC method regardless of sample size and confidence level. In comparison with their nominal coverage probabilities, they provide coverage that is at least 12% lower than the nominal coverage probability of 0.95, and they are at least 14.5% below the 0.90 nominal coverage probability. For LC prediction intervals, a predictive bootstrap method, such as PB-B and NPI-B, only performs well and produces good coverage.

Also, we compare the performance of different bootstrap methods using the MT prediction interval. PP-B and NPI-B have over-coverage for all cases of the MT method arising from a larger average width of the prediction intervals. PB and EB produce coverage close to the advertised levels of confidence with the MT method.

Bootstrap	measures	Confidence level							
		95%				90%			
		$n=50$	$n=100$	$n=200$	$n=400$	$n=50$	$n=100$	$n=200$	$n=400$
PP-B	CP_{LC}	0.9330	0.9400	0.9450	0.9410	0.8720	0.8850	0.8770	0.8900
	CP_{MT}	0.9350	0.9410	0.9470	0.9410	0.8760	0.8870	0.8780	0.8900
	AW_{LC}	3.1217	3.1253	3.1326	3.1339	2.6093	2.6147	2.6184	2.6212
	AW_{MT}	3.1527	3.1415	3.1408	3.1380	2.6349	2.6281	2.6250	2.6245
NPI-B	CP_{LC}	0.9510	0.9460	0.9490	0.9410	0.8960	0.8980	0.8890	0.9000
	CP_{MT}	0.9520	0.9480	0.9490	0.9420	0.9000	0.8980	0.8900	0.9000
	AW_{LC}	3.4439	3.2756	3.2166	3.1693	2.7613	2.6787	2.6528	2.6394
	AW_{MT}	3.4899	3.2954	3.2260	3.1733	2.7932	2.6941	2.6598	2.6428
PB	CP_{LC}	0.9330	0.9400	0.9450	0.9410	0.8720	0.8850	0.8770	0.8900
	CP_{MT}	0.9350	0.9410	0.9470	0.9410	0.8760	0.8870	0.8780	0.8900
	AW_{LC}	3.1217	3.1253	3.1326	3.1339	2.6093	2.6147	2.6184	2.6212
	AW_{MT}	3.1527	3.1415	3.1408	3.1380	2.6349	2.6281	2.6250	2.6245
EB	CP_{LC}	0.9220	0.9350	0.9430	0.9420	0.8730	0.8830	0.8810	0.8950
	CP_{MT}	0.9280	0.9360	0.9440	0.9420	0.8770	0.8840	0.8810	0.8970
	AW_{LC}	3.0576	3.1101	3.1305	3.1278	2.6044	2.6029	2.6138	2.6206
	AW_{MT}	3.0928	3.1249	3.1389	3.1320	2.6242	2.6163	2.6217	2.6239

Table 3.12: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for a single future observations from Gamma(6,3), when* $m=1$.

It is possible that the coverage of PP-B and NPI-B with the LC method is closer to the nominal coverage probability than PB and EB with the MT method. For example, PP-B and NPI-B have closer coverage with the LC method to nominal coverage probabilities for $m = 400$ than PB and EB with the MT method.

A single future observation is considered to study the coverage proportion and average interval width of the percentile prediction interval based on LC and MT methods using the same past sample sizes and confidence levels. The LC and MT methods achieve good coverage in all cases of a signal future observation with different bootstrap methods as shown in Table 3.12. A smaller sample produces wider intervals, leading to the single future observation having a greater average width of intervals compared to the mean of m future observations in Table 3.11. The PP-B and PB methods have exactly the same coverage and average interval widths in all cases of the LC and MT methods. The reason for this is that both methods have the same bootstrap sample as they consider a single observation in each bootstrap sample, also the same seeds are used for generating different bootstrap methods.

3.5 Percentile confidence interval

The BCa method was used to compare different bootstrap approaches through simulations in order to evaluate their performance in estimation. The comparison was conducted using the Beta(8,2) and N(4,3) distributions to simulate data. The greater variability in the bootstrap samples of PP-B and NPI-B caused wider intervals than other bootstrap methods. A wide interval leads to over-coverage results in all cases when using NPI-B. It is surprising that PP-B shows sometimes under-coverage results despite having wide intervals. Efron [40] recommends the BCa interval method for general use, in particular for nonparametric problems. The PB method performed poorly with under-coverage for all cases of the median with Beta(8,2) and N(3,4). The PP-B method achieves much better median coverage than the PB method, but it still shows under-coverage despite its wide intervals. The under-coverage occurs due to the BCa method tending to produce large values of bias-correction when PP-B and PB are used for the median. The endpoints of the BCa interval are determined by the percentiles of the bootstrap distribution depending on bias-correction and acceleration values. The large bias-correction values strongly influence the BCa interval endpoints in Equation (2.27).

We study PP-B and PB with the BC interval, which is a special case of the BCa method when the acceleration value is zero. The aim of this study is to discover if the bias-correction value is indeed responsible for under-coverage of PP-B and PB for the median with Beta(8,2) and N(3,4). Table 3.13 presents the result of coverage and average interval widths using PP-B and PB with BC method. It should be noted that the median results for PP-B and PB using the BC method are exactly the same as those using the BCa method in Tables 3.5 and 3.10. The acceleration value for

(a) Beta(8,2)

Bootstrap	measures	Confidence level							
		95%				90%			
		$n=50$	$n=100$	$n=200$	$n=400$	$n=50$	$n=100$	$n=200$	$n=400$
PP-B	CP	0.9260	0.9210	0.9320	0.9200	0.8760	0.8680	0.8740	0.8560
	AW	0.1075	0.0768	0.0541	0.0383	0.0909	0.0649	0.0458	0.0324
PB	CP	0.8350	0.8330	0.8500	0.8220	0.7590	0.7590	0.7700	0.7420
	AW	0.0806	0.0579	0.0411	0.0290	0.0687	0.0493	0.0349	0.0246

(b) N(3,4)

Bootstrap	measures	Confidence level							
		95%				90%			
		$n=50$	$n=100$	$n=200$	$n=400$	$n=50$	$n=100$	$n=200$	$n=400$
PP-B	CP	0.9170	0.9110	0.9240	0.9150	0.8580	0.8590	0.8430	0.8450
	AW	1.7279	1.2230	0.8627	0.6109	1.4576	1.0367	0.7309	0.5168
PB	CP	0.8320	0.8230	0.8090	0.8140	0.7580	0.7310	0.7270	0.7440
	AW	1.2991	0.9182	0.6526	0.4642	1.1061	0.7835	0.5559	0.3946

Table 3.13: *Coverage of* $100(1-2\alpha)\%$ *confidence interval using BC method for the median with original sample from Beta(8,2) and N(3,4).*

the median is zero, and therefore the endpoints of the BCa and BC intervals are the same as discussed in Section 2.5. We also use the same original sample sizes for both interval methods and the same seeds are applied for all bootstrap methods. It is important to note that the under-coverage for PP-B and PB with BCa method is not only restricted to the median. In Appendix A, we present the simulation results for the BCa method with Exp(4). The results show that PP-B and PB have worse undercoverage results for all cases of variance, as well as for the median. The BC interval was also studied in Appendix A using PP-B and PB with Exp(4) in order to determine whether bias-correction is the cause of undercoverage. The BC methods applied to PP-B and PB are under-coverage for all cases of variance and median. Note, the acceleration value for the variance is not zero, and therefore the endpoints of the BCa and BC intervals are not the same.

Chernick and LaBudde [9] discuss bootstrap confidence intervals for estimating variance in a nonparametric setting. They examine the coverage of different confidence intervals using the EB method for different sample sizes and distributions. Through extensive simulations, they have shown that the percentile method performs nearly and sometimes better than the BCa method. The percentile interval endpoints are given by the percentiles of the bootstrap distribution directly. The BCa interval endpoints are also determined by the percentiles of the bootstrap distribution, but not necessarily the same ones as those for the percentile interval. The percentiles used for the BCa interval are based on two values: acceleration and bias-correction. The BCa interval provides the same estimate of the percentile interval when both acceleration and bias-correction are equal to zero. We use the percentile interval in Equation (2.17) to study the performance of different bootstrap methods and compare it with the BCa interval. Simulations are conducted using the same original samples of Beta(8,2) and N(3,4) that were considered in the BCa interval studies. Tables 3.14 and 3.15 show the results of coverage and average interval widths for several statistics using Beta(8,2) and N(3,4), respectively.

There is a clear difference between the two interval methods with PP-B and PB for the median of both distributions. The BCa method shows under-coverage results for all cases of the median with Beta (8,2) and N(3,4). Conversely, PP-B and PB have over-coverage results for all cases of the median in the percentile method with Beta (8,2) and N(3,4). The percentile method produced similar results as the BCa method for the mean and variance with both distributions, also for the median when using NPI-B and EB. This occurs because the BCa method has smaller values of bias-correction and acceleration, resulting in the endpoints of the BCa interval being close to the percentile interval. Additionally, the simulation results for the percentile method with Exp(4) are presented in Appendix A. The results show that PP-B and PB have over-coverage for all cases of variance and median in contrast to the BCa method.

(a) mean

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9850	0.9910	0.9930	0.9910	0.9670	0.9750	0.9820	0.9700
	AW	0.0933	0.0665	0.0471	0.0333	0.0780	0.0558	0.0396	0.0280
NPI-B	CP	0.9970	0.9970	0.9950	0.9930	0.9860	0.9850	0.9840	0.9690
	AW	0.1089	0.0724	0.0492	0.0341	0.0901	0.0605	0.0412	0.0286
PB	CP	0.9430	0.9410	0.9540	0.9440	0.8800	0.8980	0.9080	0.9050
	AW	0.0664	0.0472	0.0334	0.0236	0.0558	0.0397	0.0281	0.0198
EB	CP	0.9350	0.9430	0.9580	0.9460	0.8800	0.9030	0.9090	0.9040
	AW	0.0656	0.0469	0.0332	0.0235	0.0552	0.0395	0.0279	0.0198

(b) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9820	0.9900	0.9930	0.9940	0.9590	0.9770	0.9800	0.9740
	AW	0.0175	0.0126	0.0090	0.0063	0.0144	0.0105	0.0075	0.0053
NPI-B	CP	0.9980	0.9960	0.9960	0.9950	0.9920	0.9860	0.9870	0.9870
	AW	0.0390	0.0227	0.0134	0.0082	0.0308	0.0181	0.0108	0.0067
PB	CP	0.9450	0.9410	0.9520	0.9520	0.9000	0.9020	0.9020	0.9000
	AW	0.0128	0.0091	0.0064	0.0045	0.0108	0.0076	0.0054	0.0038
EB	CP	0.8820	0.9320	0.9380	0.9450	0.8310	0.8750	0.8880	0.8870
	AW	0.0117	0.0087	0.0063	0.0045	0.0099	0.0074	0.0053	0.0037

(c) median

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9930	0.9990	0.9950	0.9980	0.9850	0.9870	0.9910	0.9840
	AW	0.1086	0.0777	0.0550	0.0389	0.0911	0.0652	0.0462	0.0327
NPI-B	CP	0.9910	0.9900	0.9880	0.9910	0.9700	0.9730	0.9720	0.9810
	AW	0.1210	0.0857	0.0608	0.0427	0.1017	0.0719	0.0512	0.0359
PB	CP	0.9770	0.9830	0.9810	0.9770	0.9430	0.9510	0.9620	0.9430
	AW	0.0838	0.0600	0.0426	0.0301	0.0705	0.0505	0.0358	0.0253
EB	CP	0.9470	0.9380	0.9480	0.9420	0.8990	0.8860	0.8950	0.8830

(a) mean

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9960	0.9960	0.9940	0.9940	0.9800	0.9830	0.9770	0.9790
	AW	1.5543	1.1027	0.7798	0.5511	1.2996	0.9240	0.6545	0.4632
NPI-B	*CP*	0.9970	0.9990	0.9960	0.9950	0.9860	0.9890	0.9800	0.9850
	AW	1.6698	1.1508	0.7997	0.5599	1.3932	0.9634	0.6709	0.47696
PB	CP	0.9530	0.9560	0.9500	0.9440	0.9080	0.9030	0.9070	0.8930
	AW	1.1058	0.7808	0.5524	0.3900	0.9295	0.6566	0.4641	0.3281
EB	CP	0.9520	0.9520	0.9560	0.9470	0.9090	0.9020	0.9020	0.8960
	AW	1.0940	0.7778	0.5510	0.3899	0.9199	0.6548	0.4633	0.3276

(b) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9850	0.9910	0.9930	0.9920	0.9620	0.9780	0.9770	0.9730
	AW	4.3649	3.1047	2.2035	1.5569	3.6152	2.5903	1.8443	1.3061
NPI-B	*CP*	0.9960	0.9910	0.9940	0.9950	0.9810	0.9800	0.9800	0.9850
	AW	5.7466	3.8038	2.5375	1.7074	4.6725	3.1211	2.1020	1.4230
PB	CP	0.9410	0.9520	0.9480	0.9480	0.8910	0.9060	0.8950	0.8970
	AW	3.1864	2.2303	1.5700	1.1051	2.6754	1.8728	1.3207	0.9285
EB	CP	0.9170	0.9400	0.9410	0.9410	0.8540	0.8990	0.8840	0.8900
	AW	2.9811	2.1597	1.5450	1.0950	2.5092	1.8171	1.2997	0.9206

(c) median

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	1.0000	1.0000	0.9980	0.9970	0.9920	0.9950	0.9910	0.9940
	AW	1.7632	1.2504	0.8842	0.6249	1.4725	1.0476	0.7424	0.5249
NPI-B	*CP*	0.9940	0.9920	0.9930	0.9960	0.9770	0.9770	0.9750	0.9820
	AW	1.9760	1.3841	0.9795	0.6917	1.6521	1.1602	0.8248	0.5808
PB	CP	0.9880	0.9900	0.9860	0.9850	0.9610	0.9650	0.9610	0.9610
	AW	1.3678	0.9728	0.6892	0.4877	1.1489	0.8183	0.5796	0.4103
EB	CP	0.9520	0.9420	0.9430	0.9480	0.9100	0.9040	0.8910	0.8890

3.6 Prediction interval for a future sample statistic

The prediction performance of different bootstrap methods was compared using prediction intervals for the mean of m future observations based on LC and MT methods. The bootstrap methods with a predictive nature, that is PP-B and NPI-B, only performed well with the LC method. The MT method for the mean of m future observations improves coverage in PB and EB because they approximate the nominal coverage probabilities. In the case of a prediction for $m > 1$ future observations, we use the mean of m future observations to compute the percentile prediction interval for future sample mean. The percentile prediction interval based on the bootstrap method can be generalized to a large class of statistics and is not restricted to sample means. In this section, we investigate the prediction intervals for a statistic of m future observations based on LC and MT methods.

A simulation study is conducted in the same manner as described in Section 3.3.3, except that the statistic of m future observations (T_m) is computed rather than the future sample mean ($\bar{y}_m$). We set the number of simulations at $N = 1000$ and the bootstrap methods are applied to each past sample $B = 1000$ times. Mojsheibani [59] investigated the prediction intervals using a future sample m with a different size from the past sample n. We construct the percentile prediction intervals for the variance of $m = n/2$ future observations. Simulations are conducted with the variance statistic by applying the same distributions, past samples, and confidence levels that were used in the future sample mean studies. We use different original sample sizes $n = 50, 100, 200, 400$ from Beta(3,1) and Gamma(6,3) with confidence levels 95% and 90%. The results of coverage proportions and interval average widths of the future sample variance using LC and MT methods for Beta(3,1) and Gamma(6,3) are presented in Tables 3.16 and 3.17, respectively.

The performance of different bootstrap procedures is first compared with the LC prediction interval. PP-B and NPI-B have good coverage in all cases of future sample sizes $m = n/2$ at confidence levels 95% and 90%. The superiority of PP-B is that it achieves good coverage with shorter intervals, where the average interval widths

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP_{LC}	0.9510	0.9460	0.9490	0.9410	0.8990	0.8800	0.9020	0.8760
	CP_{MT}	0.9800	0.9770	0.9810	0.9760	0.9560	0.9500	0.9550	0.9490
	AW_{LC}	0.0506	0.0361	0.0258	0.0183	0.0425	0.0303	0.0217	0.0154
	AW_{MT}	0.0618	0.0440	0.0315	0.0223	0.0520	0.0371	0.0265	0.0188
NPI-B	CP_{LC}	0.9660	0.9610	0.9570	0.9530	0.9320	0.9230	0.9070	0.8950
	CP_{MT}	0.9960	0.9880	0.9890	0.9880	0.9690	0.9660	0.9620	0.9570
	AW_{LC}	0.0629	0.0416	0.0279	0.0191	0.0522	0.0346	0.0233	0.0160
	AW_{MT}	0.0782	0.0514	0.0343	0.0233	0.0648	0.0428	0.0287	0.0196
PB	CP_{LC}	0.8110	0.8430	0.8310	0.8590	0.7250	0.7650	0.7600	0.7860
	CP_{MT}	0.8920	0.9060	0.9050	0.9240	0.8230	0.8520	0.8440	0.8730
	AW_{LC}	0.0417	0.0299	0.0211	0.0150	0.0353	0.0252	0.0178	0.0126
	AW_{MT}	0.0504	0.0363	0.0257	0.0183	0.0428	0.0307	0.0217	0.0154
EB	CP_{LC}	0.8700	0.8620	0.8800	0.8560	0.8050	0.7980	0.8050	0.7850
	CP_{MT}	0.9230	0.9170	0.9390	0.9350	0.8780	0.8750	0.8940	0.8710
	AW_{LC}	0.0405	0.0293	0.0210	0.0149	0.0344	0.0248	0.0177	0.0125
	AW_{MT}	0.0485	0.0355	0.0255	0.0181	0.0415	0.0301	0.0216	0.0153

Table 3.16: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for the variance of* m *future observations from Beta(3,1), when* $m = n/2$.

of PP-B are smaller than NPI-B in all cases. Additionally, its coverage proportions are closer to the coverage probabilities than those of NPI-B in most cases. In contrast, PB and EB show worse under-coverage results for all cases of Beta(3,1) and Gamma(6,2). Their coverage proportions with Beta(3,1) are at least 7% lower than 0.95 and 0.90 nominal coverage probabilities. Also, they provide coverage proportions that are at least 5.2% below their nominal coverage probabilities with Gamma(6,3).

The performance of different bootstrap methods is also compared based on MT prediction intervals. The wide average width of intervals in both PP-B and NPI-B leads to over-coverage for all cases. The MT method improves the coverage

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP_{LC}	0.9300	0.9480	0.9450	0.9420	0.8670	0.8910	0.8980	0.8830
	CP_{MT}	0.9750	0.9840	0.9840	0.9770	0.9390	0.9530	0.9530	0.9490
	AW_{LC}	1.0694	0.7646	0.5456	0.3875	0.8631	0.6272	0.4520	0.3228
	AW_{MT}	1.3956	0.9715	0.6804	0.4787	1.1071	0.7890	0.5621	0.3989
NPI-B	CP_{LC}	0.9630	0.9750	0.9650	0.9660	0.9200	0.9410	0.9310	0.9250
	CP_{MT}	0.9900	0.9930	0.9960	0.9900	0.9680	0.9780	0.9720	0.9700
	AW_{LC}	2.2296	1.3233	0.8047	0.5070	1.5779	0.9800	0.6180	0.4025
	AW_{MT}	3.6831	2.0642	1.1798	0.6992	2.3717	1.3936	0.8430	0.5272
PB	CP_{LC}	0.8810	0.8980	0.8920	0.8800	0.8180	0.8320	0.8320	0.8110
	CP_{MT}	0.9510	0.9560	0.9480	0.9440	0.8920	0.9090	0.9030	0.8890
	AW_{LC}	0.9109	0.6378	0.4500	0.3176	0.7517	0.5302	0.3755	0.2664
	AW_{MT}	1.1492	0.7909	0.5526	0.3890	0.9399	0.6564	0.4630	0.3264
EB	CP_{LC}	0.8310	0.8640	0.8730	0.8730	0.7550	0.8080	0.8090	0.8010
	CP_{MT}	0.8960	0.9190	0.9200	0.9340	0.8340	0.8760	0.8790	0.8860
	AW_{LC}	0.7936	0.5934	0.4317	0.3102	0.6773	0.5020	0.3631	0.2612
	AW_{MT}	0.9565	0.7193	0.5244	0.3778	0.8137	0.6092	0.4431	0.3186

Table 3.17: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for the variance of* m *future observations from Gamma(6,3), when* $m = n/2$.

proportions of PB and EB, but PB is at least 4.4% below their nominal coverage probabilities with Beta(3,1) when $n = 50, 100, 200$ as shown in Table 3.16. Also, the EB method gives a result of 5.4% under-coverage below the nominal level of 95% and 6.6% lower than the nominal level of 90% with Gamma(6,3) when $n = 50$ as shown in Table 3.17. We observe that the MT method improves the coverage probability of PB and EB, however we can obtain a coverage proportion that is close to the nominal coverage probability using PP-B and NPI-B with the LC method. For example, the coverage of PP-B and NPI-B with Beta(3,1) based on LC method is closer to the nominal coverage probabilities when $n = 50$ than PB and EB with MT method. The MT method enhances the coverage proportions of PB and EB by expanding the prediction interval width, but it provides under-coverage results in some cases. It is obvious that the PP-B method performs best for LC prediction intervals, as it is developed for predictive inference in line with the NPI-B method.

3.7 Concluding remarks

This chapter introduced a new bootstrap method, the parametric predictive bootstrap (PP-B). It has been applied to a variety of scenarios that have been used with other bootstrap methods, to investigate its performance in estimation and prediction. First, we studied the estimation performance of different bootstrap methods using some measures of statistical accuracy: standard error, bias, root mean squared error, and mean absolute error. The smaller values of statistical accuracy measures are considered to be good characteristics of an estimator. A sampling method for PP-B and NPI-B that involves adding the sampled observation to the data before sampling the next observation, thereby increasing the variation of their bootstrap samples. Consequently, the values of statistical accuracy measures are generally greater in PP-B and NPI-B than in the other bootstrap methods.

Confidence intervals are used to compare the performance of different bootstrap methods as an estimation approach. We consider the BCa and percentile methods to construct confidence intervals based on a bootstrap technique for several statistics: mean, variance and median. The PP-B and NPI-B methods have an over-coverage tendency due to wider intervals arising from greater variability in their bootstrap samples. However, the over-coverage of PP-B disappears for the median when the BCa interval is applied to all distributions that were considered in the studies, as well as the variance associated with Exp(4). The PB method also shows worse under-coverage in the same situations as the PP-B method which shows under-coverage with a BCa interval. The BCa interval is specifically designed for nonparametric problems, and it depends on bias-correction and acceleration values. We notice that the bias-correction values are large when we use the PP-B and PB methods, in particular for the median. The BCa interval endpoints have been adversely affected by these large values. It is not surprising that PP-B does not provide confidence intervals with the right coverage, as it is developed for predictive inference.

Finally, we evaluated the prediction performance of different bootstrap methods by considering two prediction intervals: the LC and MT methods. A proportion of coverage close to nominal coverage probability is desirable, along with a shorter average interval width. The method for sampling observations in PP-B leads to variation in different bootstrap samples, reflecting uncertainty regarding the predictions consistent with NPI-B. Predictive bootstrapping methods, such as PB-B and NPI-B, perform well for the LC method and lead to giving good coverage. A major advantage of PP-B is that it provides good coverage with a shorter average width of intervals than NPI-B, as shown in all simulation studies we have conducted. Conversely, PB and EB perform poorly and produce under-coverage results that are far from nominal coverage probabilities in all cases. The MT prediction interval enhances the coverage proportions of PB and EB by expanding the interval width. However, they do not perform well in some cases for the variance of $m = n/2$ future observations with the MT method.

Chapter 4

Reproducibility using Bootstrap

4.1 Introduction

The reproducibility of test outcomes is an important characteristic of practical statistics. The reproducibility probability (RP) has gained considerable attention in the literature, with some contributions indicating that the definition and interpretation of RP are not uniquely determined in classical frequentist statistics. In Section 2.3, we discussed the definitions of RP and some different methods to estimate it. Here, we consider the basic idea of RP, which is that the probability of the event that, if the experiment were repeated in the same way as the original experiment, would lead to the same test outcome. We regard assessing test reproducibility as a problem to be solved by predictive inference.

In this chapter, we introduce the PP-B method for the reproducibility of some parametric tests. We also compare this approach with a similar predictive bootstrap method for test reproducibility, NPI-B. The explicitly predictive nature of PP-B and NPI-B which consider future observations provides a natural formulation of inferences on RP. Test reproducibility is naturally viewed as a prediction problem rather than an estimation problem, which is well aligned with these approaches. PP-B-RP and NPI-B-RP are acronyms for the reproducibility value based on PP-B and NPI-B methods, respectively. It is important to emphasize that we primarily focus on the conclusion of the future test with respect to the null hypothesis based on the actual data of the first test. We do not consider an exact repetition in terms of the same value of the test statistic or the actual observations, nor do we rely only on the result of the first test that the null hypothesis was rejected or not. Inferring the reproducibility of the test result using actual data seems logical because the strength of the first test's conclusion depends on those data. A prediction of the test result in a future test is more naturally reflected in the final conclusion in terms of rejection or non-rejection of the null hypothesis. We should point out that the bootstrap approaches do not require that sample sizes be the same for actual and future tests, but this assumption is natural for reproducibility.

This chapter is organized as follows: A brief review of basic parametric tests is provided in Section 4.2. The PP-B method is employed for test reproducibility by considering these parametric tests in Sections 4.3, 4.4, and 4.5. We also compare this approach with a similar predictive bootstrap method for test reproducibility with those parametric tests, the NPI-B. In Section 4.6, we provide a comparison between the Bootstrap-RP and NPI-RP methods for the likelihood ratio test. The final section of this chapter ends with some concluding remarks.

4.2 Overview of some parametric tests

In this section we provide an overview of some parametric tests: The one-sample t-test, two-sample t-test, Welch's t-test, and F-test. Reproducibility of these tests will be considered later in this chapter. A statistical hypothesis test is a method of statistical inference employed by the analyst to evaluate the plausibility of a hypothesis about the population using sample data. Typically, a hypothesis test involves a pair of opposing hypotheses. The null hypothesis, H_0, assumes that any kind of difference between the chosen characteristics that you see in a set of data is due to chance. The alternative hypothesis, H_a, is contradictory to the null hypothesis which states that a difference in the population characteristic is not due to a chance occurrence. Researchers test for a significant effect by computing the probability of rejecting H_0 with the value of the test statistic under the assumption that the null hypothesis is correct. The computed probability is known as p-value or attained significance level, if it is less than a predetermined level of significance, then the null hypothesis is rejected [58].

In the hypothesis testing procedure, the significance level α has a predetermined value typically chosen to be 0.05 or 0.10. The p-value is a different approach from the critical value, but they lead to the same conclusion in terms of rejection or non-rejection of the null hypothesis. The probability of rejecting a true null hypothesis is a type I error, and the probability of non-rejection of a false null hypothesis is a type II error. These error probabilities are denoted by α and β, respectively. There is an inverse relationship between Type I and Type II errors, meaning that when one increases the other decreases. The power is the probability of rejecting the false null hypothesis when a specific alternative hypothesis is true. It can be thought of as the probability of making a correct decision about the false null hypothesis, this probability is $1 - \beta$. The power increases with the α level associated with the hypothesis testing procedure and with the sample size of the experiment [58].

4.2.1 One-sample t-test

The one-sample t-test is a statistical test that can be used when comparing the mean of one group to a value. The purpose of the test is to determine whether an unknown population mean differs from a specific value. The test statistic of t-test follows a t-distribution, also known as Student's t-distribution due to William Gosset who developed and published the distribution under the pseudonym "Student" [70]. The one-sample t-test can only be applied to data that follow a normal distribution [41]. Suppose $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ is random sample from normal population with unknown variance σ^2. Generally, the hypotheses of interest are

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_a : \mu \neq \mu_0,\ \mu > \mu_0,\ \mu < \mu_0$$

Under the assumption that the one sample comes from a normal distribution, the test statistic can be computed under the null hypothesis by the following formula:

$$T = \frac{\bar{x} - \mu}{\sqrt{s^2/n}} \sim t_{n-1} \tag{4.1}$$

where t_{n-1} represents the Student's t-distribution with $n - 1$ degree of freedom and $\bar{x}$, s^2 are the mean and variance of sample X.

The null hypothesis H_0 is rejected in favour of the two sided test $H_a : \mu \neq \mu_0$ with the level of significant α if and only if $|T| > t_{n-1}^{(1-\alpha/2)}$, where $t_{n-1}^{(1-\alpha/2)}$ is the $(1-\alpha/2)$th percentile of the Student's t-distribution with $n-1$ degree of freedom. For one sided upper tail test $H_a : \mu > \mu_0$, we rejects H_0 if $T > t_{n-1}^{(1-\alpha)}$, and reject H_0 if $T < t_{n-1}^{(\alpha)}$ for one sided lower tail test $H_a : \mu < \mu_0$.

4.2.2 Two-sample t-test

Clinical studies typically assign patients randomly to two groups: the treatment group and the control group. In the treatment group, patients receive the treatment or a test drug while patients in the control group receive the placebo. The researcher measures the patients' responses from both groups as part of evaluating the effectiveness of the treatment to investigate whether there is a significant difference in the mean values between the two groups. The two-sample t-test is used to compare the means between two groups, which is considered one of the most commonly used statistical hypothesis tests in pain studies [50]. It is frequently used in health research to determine whether a treatment actually has an effect on the population of interest or not, also to compare the differences between two groups. The two-sample t-test (also known as pooled variance t-test) can be carried out for a comparison of two means when both samples meet certain requirements of statistical assumptions as it is a parametric test. It is required that two samples are normally distributed with equal variances and independent of each other [69]. Suppose $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma^2)$, are two independent random samples from normal populations with unknown (common) variance σ^2. Generally, the hypotheses of interest are

$$H_0 : \mu_1 = \mu_2 \quad \text{versus} \quad H_a : \mu_1 \neq \mu_2,\ \mu_1 > \mu_2,\ \mu_1 < \mu_2$$

Under the assumption that the two samples are independent and come from a normal distribution with equal variances, the test statistic can be computed under the null hypothesis by the following formula:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{s_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2} \tag{4.2}$$

where t_{n+m-2} represents the Student's t-distribution with $n+m-2$ degree of freedom and $\bar{x}$, $\bar{y}$, s_1^2 , s_2^2 are the means and variances of two samples X and Y, respectively. A pooled variance of the two samples is defined as follows:

$$s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$$

For the one sided upper tail test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 > \mu_2$ with the level of significant α, we reject H_0 if $T > t_{n+m-2}^{(1-\alpha)}$, where $t_{n+m-2}^{(1-\alpha)}$ is the $(1-\alpha)$th percentile of the Student's t-distribution with $n+m-2$ degree of freedom, while reject H_0 if $T < t_{n+m-2}^{(\alpha)}$ for one sided lower tail test $H_a : \mu_1 < \mu_2$. If we use the two sided test $H_a : \mu_1 \neq \mu_2$, we reject H_0 if $|T| > t_{n+m-2}^{(1-\alpha/2)}$.

4.2.3 Welch's t-test

Welch introduced another version adapted from the Student's t-test, which can be used when there is a significant difference between the variances of the two samples [1, 67]. Welch's t-test (also known as non-pooled variance t-test and unequal variance t-test) assumes that the sample means being compared for two populations are normally distributed, but it is designed for unequal population variances. Suppose $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$, are two independent random samples from normal populations with unequal variances. Under this assumption, the test statistic can be computed under the null hypothesis by the following formula:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}} \sim t_v \tag{4.3}$$

The degrees of freedom v associated with this variance estimate can be approximated based on the sample size and variance of each sample as follows:

$$v = \frac{(s_1^2/n + s_2^2/m)^2}{\left(\frac{s_1^2}{n}\right)^2/(n-1) + \left(\frac{s_2^2}{m}\right)^2/(m-1)}$$

The Student's t-test devised by Gosset assumes variances of two samples are equal. Therefore, we estimate the common variance as a weighted average of two sample variances (pooled variance). Welch's t-test is suitable when the assumption of equal variance is not met. The degrees of freedom associated with Welch's t-test are a random variable dependent on the sample sizes and the variances of the samples. Welch's test always has degrees of freedom less than or equal to the degrees of freedom used in the Student's t-test [27]. The approximate degrees of freedom used in Welch's test are more conservative than the degrees of freedom used in the two-sample t-test. The null hypothesis H_0 is rejected in favour of the one sided upper tail test $H_a : \mu_1 > \mu_2$ with the level of significant α if $T > t_v^{(1-\alpha)}$, where $t_v^{(1-\alpha)}$ is the $(1-\alpha)$th percentile of the Student's t-distribution with v degree of freedom, and reject H_0 if $T < t_v^{(\alpha)}$ for one sided lower tail test $H_a : \mu_1 < \mu_2$. For two sided test $H_a : \mu_1 \neq \mu_2$, We rejects H_0 if and only if $|T| > t_v^{(1-\alpha/2)}$.

4.2.4 F-test

The F-test of equality of variances is a test for the null hypothesis that the variances of two normal samples are the same. It is known as the F-ratio test due to the way of computing test statistics by the ratio of two sample variances. The F-test is valid for equality of variances under the assumption of normality for two samples and when this assumption is in doubt we should use alternative test to compare variances between two samples [48]. However, the two-sample t-test and Welch's t-test required normality assumption of two samples to be performed which is agreed with the F-test. Suppose $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$, be two independent random samples from normal populations. Generally, we test the hypotheses:

$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_a : \sigma_1^2 \neq \sigma_2^2,\ \sigma_1^2 > \sigma_2^2,\ \sigma_1^2 < \sigma_2^2$$

The F-test is a test for the null hypothesis that the two samples from normal distributions have the same variance and the test statistic can be computed as the ratio of the two variances by the following formula:

$$F = \frac{s_1^2}{s_2^2} \sim F_{n-1,m-1} \tag{4.4}$$

where $F_{n-1,m-1}$ represents the F distribution with $n-1$ and $m-1$ degree of freedom. The test statistics of F-test is greater than or equal to zero. The decision rule for two sided test is to reject H_0 at significance level α if and only if $F < F^{(\alpha/2)}_{n-1,m-1}$ or $F > F^{(1-\alpha/2)}_{n-1,m-1}$, where $F^{(\alpha/2)}_{n-1,m-1}$ is $(\alpha/2)$th percentile of the F-distribution with $n-1$ and $m-1$ degrees of freedom. The null hypothesis H_0 is rejected in favour of the one sided upper tail test $H_a : \sigma_1^2 > \sigma_2^2$ with the level of significant α if $F > F^{(1-\alpha)}_{n-1,m-1}$, and reject H_0 if $F < F^{(\alpha)}_{n-1,m-1}$ for one sided lower tail test $H_a : \sigma_1^2 < \sigma_2^2$.

4.3 Bootstrap-RP for the one-sample t-test

In this section, we study the RP of one-sample t-test using the bootstrap method. The PP-B method is employed for RP by considering the one-sample t-test and comparing its performance with a similar predictive bootstrap method for RP, the NPI-B. Test reproducibility is naturally considered as a predictive inference problem and the explicitly predictive nature of PP-B and NPI-B provides an appropriate formulation for inferring RP. We provide a comparison through simulation studies to get an insight into the performance of the two bootstrap methods with the RP of one-sample t-test. These are performed as follows:

1. Apply the one-sample t-test to the original sample X of size n to obtain the value of the test statistic, then decide whether or not the null hypothesis is rejected based on this test value.

2. Draw a bootstrap sample of size n from the sample X and apply the same test to obtain the decision of this test.

3. Perform Step 2 in total B times and record the test result each time whether the null hypothesis is rejected or not.

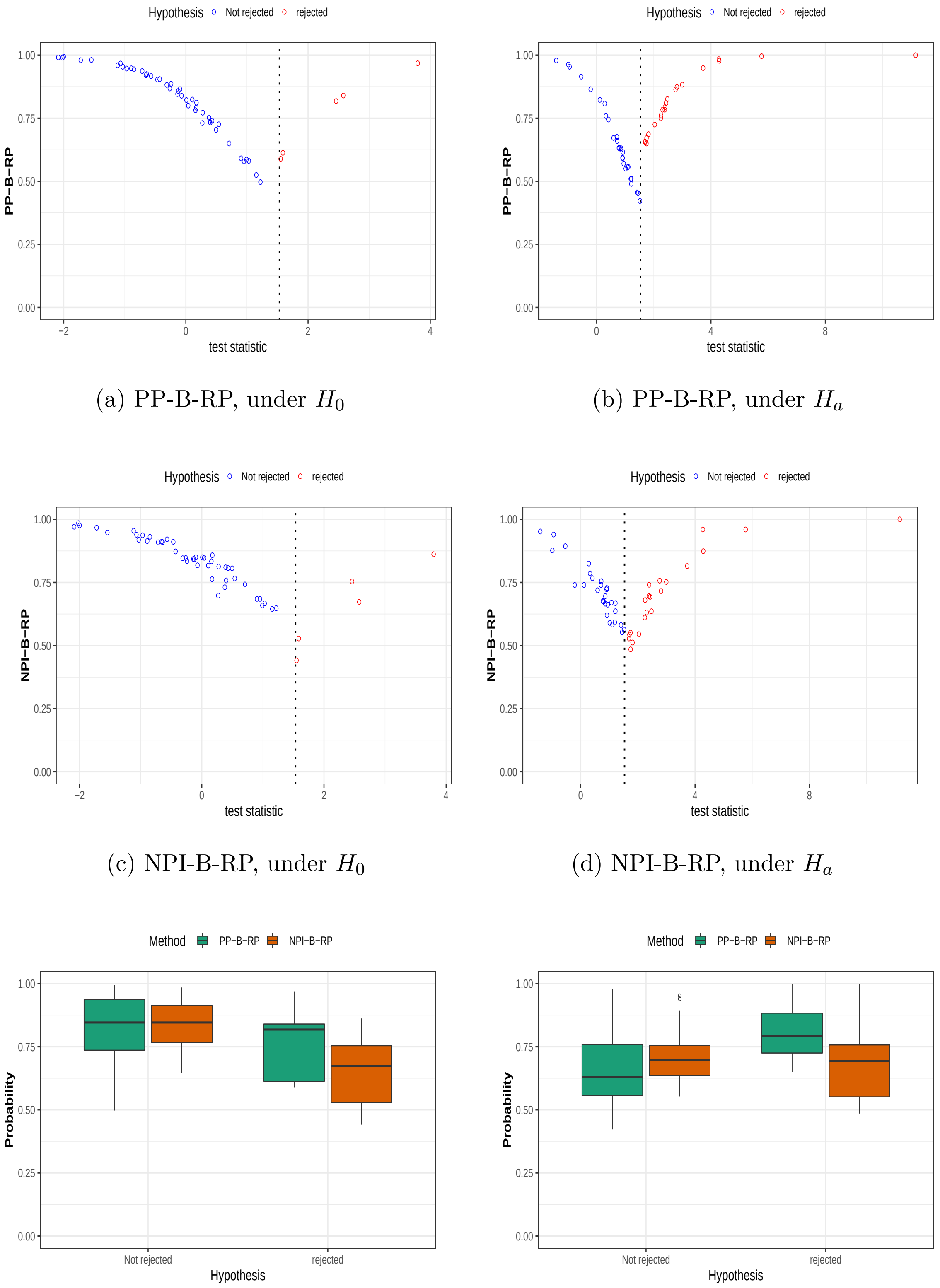
4. The estimate of the RP is the ratio of B times in which the original sample and the bootstrap samples have the same conclusion.

5. Perform all these steps N times in order to obtain RP values for both rejection and non-rejection cases of the null hypothesis.

The one sided one-sample t-test is considered, $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$, with level of significance $\alpha = 0.10$. We simulate $N = 50$ samples of size $n = 5$ under both H_0 and H_a. The data are generated from the Normal distribution with mean 0 under H_0 and mean 0.5 under H_a, both with standard deviation 1. All values of RP were determined based on the PP-B and NPI-B methods as described above using $B = 1000$ bootstrap samples. For each $N = 50$ sample, the observed test statistic and Bootstrap-RP were calculated. The same data sets for each sample are used to compute the RP value of one-sample t-test based on the two bootstrap methods. It is important to emphasize that the bootstrap samples for each method have the same size as the original sample. Figure 4.1 presents the results of RP values using the two bootstrap methods under H_0 and H_a for samples of size $n = 5$. The boxplots of RP values are displayed for the two bootstrap methods in both cases of rejection and non-rejection. The red circles in all figures in this thesis refer to rejection cases of H_0 while the blue circles represent non-rejection cases.

We first examine the relationship between Bootstrap-RP and the test statistic for one-sample t-test in the simulations. The values of RP for the two methods tend to increase when the test statistic moves away from the test thresholds, as expected, regardless of the decision on H_0. The worst-case scenario gives RP of about 0.5 when the original test statistic is close to the test threshold. In the absence of further information, one would expect a repeat experiment to produce a second test statistic whose value is equally likely to be larger or smaller than the original test statistic, and therefore the same conclusion would be reached with a probability of 0.5. A repetition of an experiment that had an original test statistic far away from the test threshold is likely to produce a second test statistic that is also far away from the test threshold. Therefore, the RP values tend to increase when the test statistic moves away from the test thresholds. Simulation studies show that RP values based on PP-B have less variability than NPI-B because of the parametric model assumed for PP-B. There is a clear fluctuation observed in the values of RP based on NPI-B because this bootstrap method does not assume a parametric model and the sample size is quite small. The fluctuation of RP values based on NPI-B is more visible when simulations are conducted under H_a due to more cases of test statistics close to the test threshold.

We also compare PP-B-RP and NPI-B-RP in both cases when the null hypothesis



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_a

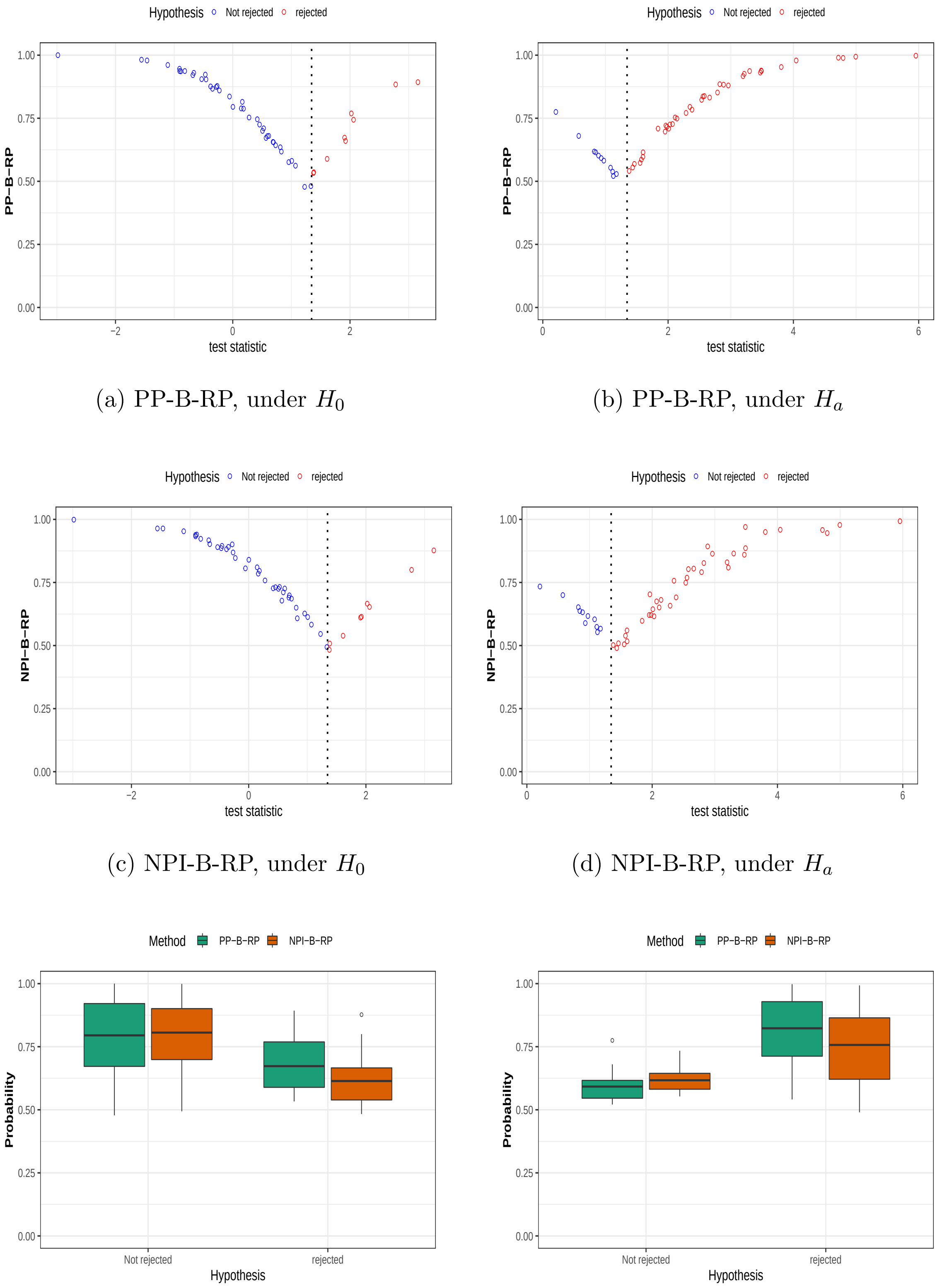
(e) RP, under H_0

(f) RP, under H_a

Figure 4.1: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for one-sample t-test, where $n = 5$.

is rejected and not rejected. It is obvious that the PP-B-RP tends to be higher in cases of rejection (red cases in the figures) than in cases of non-rejection (blue cases) when the test statistic is close to the test threshold. Conversely, NPI-B-RP tends to be lower in case of rejection than in non-rejection when the test statistic is close to the test threshold. The RP is computed by generating B bootstrap samples from the original sample and then applying the one-sample t-test for each of these bootstrap samples. Thereafter, the ratio of the B times that have the same decision as the original sample is the RP value. The test statistic of the one-sample t-test is computed using Formula (4.1), which includes the sample variance in the denominator. In general, PP-B has a smaller variance compared to NPI-B due to the assumption of a parametric model in PP-B. In the case of non-rejection, the PP-B-RP tends to be lower due to the computed test statistic from PP-B samples tending to lie in the rejection region. This occurs because PP-B samples lead to larger test statistic values than NPI-B samples, as a result of a smaller variance value in the denominator. Hence, we obtain more cases that reject H_0 due to a test statistic value being larger than the test threshold. As a result, the PP-B-RP value tends to be lower in the case of non-rejection compared to NPI-B-RP. In contrast, PP-B-RP tends to be higher in the case of rejection than NPI-B-RP. It is the same reason in the case of non-rejection, where we obtain more cases of the same decision of an original sample that does reject H_0.

Additionally, we analyze the impact of increasing sample size on the patterns of Bootstrap-RP values. The results of RP values based on the two bootstrap methods for samples of size $n = 15$ under H_0 and H_a are presented in Figure 4.2. The boxplots of RP values based on PP-B and NPI-B are shown for both rejection and non-rejection cases. As the sample size increases, the Bootstrap-RP value becomes closer to 0.5 when the observed test statistics are close to the test threshold in both cases of rejection and non-rejection. Also, the fluctuation in NPI-B-RP values is decreased when the sample size increases. The power of the test is positively correlated with sample size, which means a larger sample size gives greater power. It is because a larger sample size narrows the distribution of the test statistic, so the false null hypothesis can be distinguished more clearly



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_a

(e) RP, under H_0

(f) RP, under H_a

Figure 4.2: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for one-sample t-test, where $n = 15$.

(a) Under H_0

Sample	Test statistic	n	Test threshold	H_0	PP-B-RP	NPI-B-RP
1	1.588	5	1.533	R	0.613	0.528
2	1.551			R	0.589	0.441
3	1.221			NR	0.497	0.648
4	1.153			NR	0.525	0.645
1	1.382	15	1.345	R	0.536	0.508
2	1.377			R	0.533	0.483
3	1.333			NR	0.481	0.494
4	1.226			NR	0.478	0.546

(b) Under H_a

Sample	Test statistic	n	Test threshold	H_0	PP-B-RP	NPI-B-RP
1	1.705	5	1.533	R	0.656	0.543
2	1.689			R	0.675	0.528
3	1.516			NR	0.442	0.563
4	1.449			NR	0.453	0.553
1	1.435	15	1.345	R	0.555	0.490
2	1.378			R	0.541	0.402
3	1.176			NR	0.529	0.567
4	1.126			NR	0.521	0.553

Table 4.1: *Simulation under H_0 and H_a: values of RP of one-sample t-test using PP-B and NPI-B methods with four observed samples of sizes $n = 5$ and $n = 15$.*

from the true null hypothesis. For simulations under H_a, increasing sample size leads to more cases rejecting H_0, which simply results from the test becoming more powerful with a larger sample size. The pattern of RP values based on the two bootstrap methods changes when simulations are performed under the alternative hypothesis, resulting from changes in the observed test statistics with respect to the test threshold. Table 4.1 presents four samples close to the test threshold that reject and do not reject H_0 with sample sizes $n = 5$ and $n = 15$ for simulations under both the null and alternative hypotheses. This table includes the observed test statistics, test thresholds, PP-B-RP and NPI-B-RP. In the case of rejection, the PP-B-RP values tend to be higher than the NPI-B-RP values. Conversely, the values of PP-B-RP seem to be lower compared to the NPI-B-RP values in non-rejection cases. However, increasing n tends to reduce the differences between PP-B-RP and NPI-B-RP.

4.4 Bootstrap-RP for the two-sample t-test and Welch's t-test

In this section, we consider the RP of the two-sample t-test when both samples are normally distributed with equal variances. We also study the RP of Welch's t-test for two samples when there is a difference between the variances of the two samples. There are some technical differences and similarities between the Student's t-test and Welch's t-test. The t-values, degrees of freedom, and the p-values are the same in Student's t-test and Welch's t-test when both sample sizes and variances are the same in each sample [26]. The differences appear in both tests when variances and/or sample sizes are different. The most important difference that led to the development of Welch's t-test is when both variances and sample sizes differ between Student's t-test and Welch's t-test. In this case, the t-value, degrees of freedom, and p-value all differ in both tests due to the differences in both the variances and sample sizes. The t-value remains identical, but the degrees of freedom differ when the variance or sample size is different in both samples, so as a result, the p-value differs. Welch's t-test can be generalized to more than two samples, but we will focus on the case of the two samples [68]. Welch's t-test differs from Student's t-test in that it does not assume equal variances.

There are several tests for the assumption of equal variances such as Levene's test, the F-test, Bartlett test and Box-Andersen test [52, 53]. The F-test of equality of variances is commonly used by researchers because it is available in popular statistical software. The two-sample t-test is considered here for equal sample sizes. The simulation studies are conducted to evaluate the performance of the two bootstrap methods for RP of the two-sample t-tests through the following steps:

1. Apply the t-test on the two original samples with equal sample sizes n, X and Y, to obtain the value of the test statistic, then drew a conclusion about the null hypothesis for this test whether it is rejected or not.

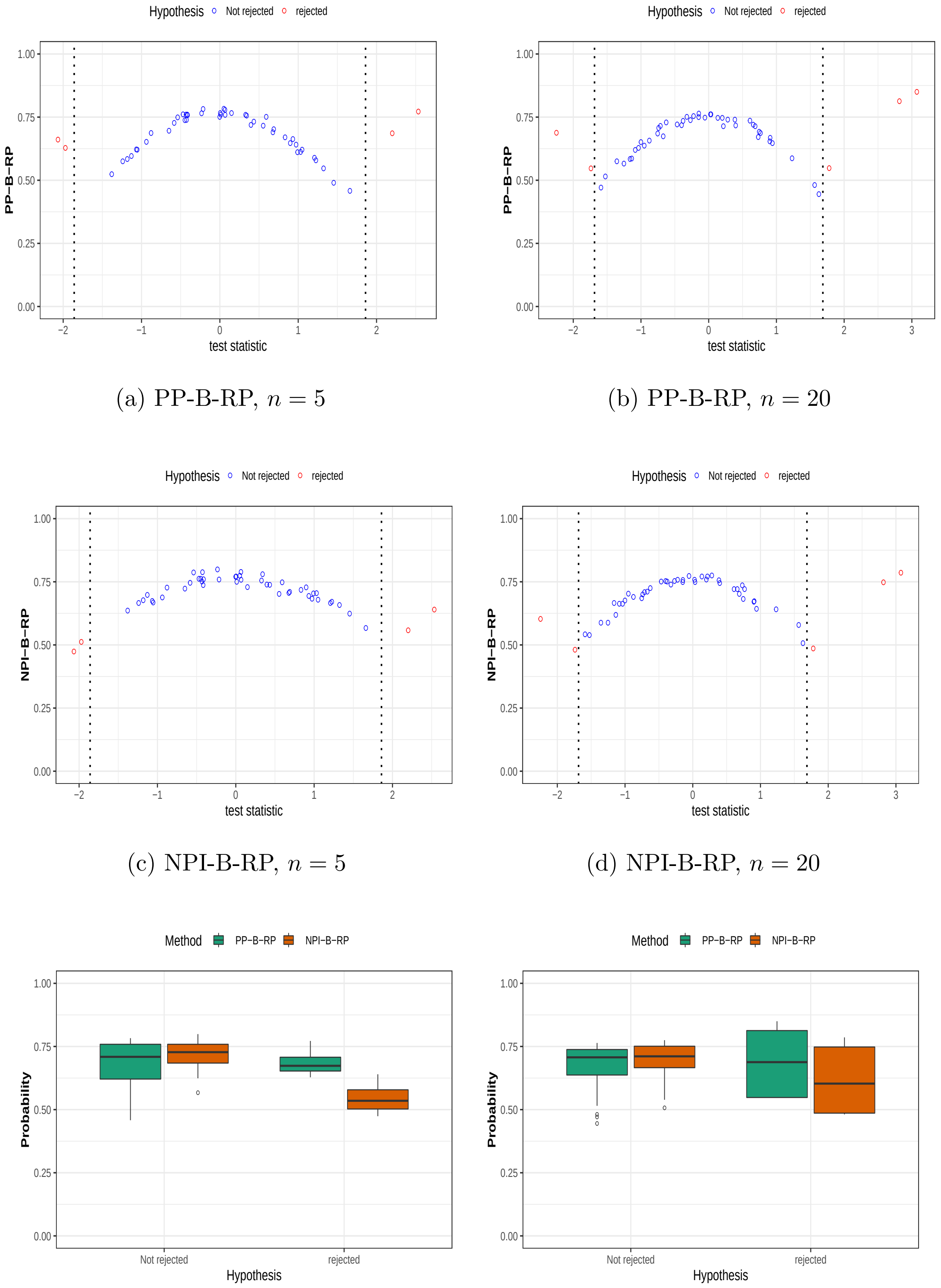
2. Draw a bootstrap sample of size n from sample X and a bootstrap sample of size n from sample Y. Apply the two-sample t-test to these two bootstrapped samples to obtain the test conclusion.

3. Perform Step 2 in total B times and record the test outcome each time whether or not the null hypothesis is rejected.

4. The ratio of B times that the two original samples and these two bootstrap samples have the same conclusion is the estimate of the RP.

5. Perform all these steps N times in order to obtain RP values for both rejection and non-rejection cases of the null hypothesis.

We first investigate the RP for the two-sample t-test when the variances of the two normally distributed populations are assumed to be equal. The two sided two-sample t-test is considered, $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$, and level of significance $\alpha = 0.10$. We simulate two samples of size $n = 5$ under H_0 in total $N = 50$ times. The data are generated for the two original samples from the same Normal distribution with mean 2 and standard deviation 1. The RP value for the two-sample t-test is computed based on the two bootstrap methods as demonstrated above using $B = 1000$ bootstrap samples. The observed test statistic and Bootstrap-RP were determined for each of $N = 50$ samples. Also, we study the impact of increasing sample size to $n = 20$ on Bootstrap-RP values for the two-sample t-test. It is important to emphasize that the same data sets are used to compute the RP values for the two-sample t-test based on PP-B and NPI-B. The results of RP values based on the PP-B and NPI-B methods with samples of size $n = 5, 20$ under H_0 are



(a) PP-B-RP, $n = 5$

(b) PP-B-RP, $n = 20$

(c) NPI-B-RP, $n = 5$

(d) NPI-B-RP, $n = 20$

(e) RP, $n = 5$

(f) RP, $n = 20$

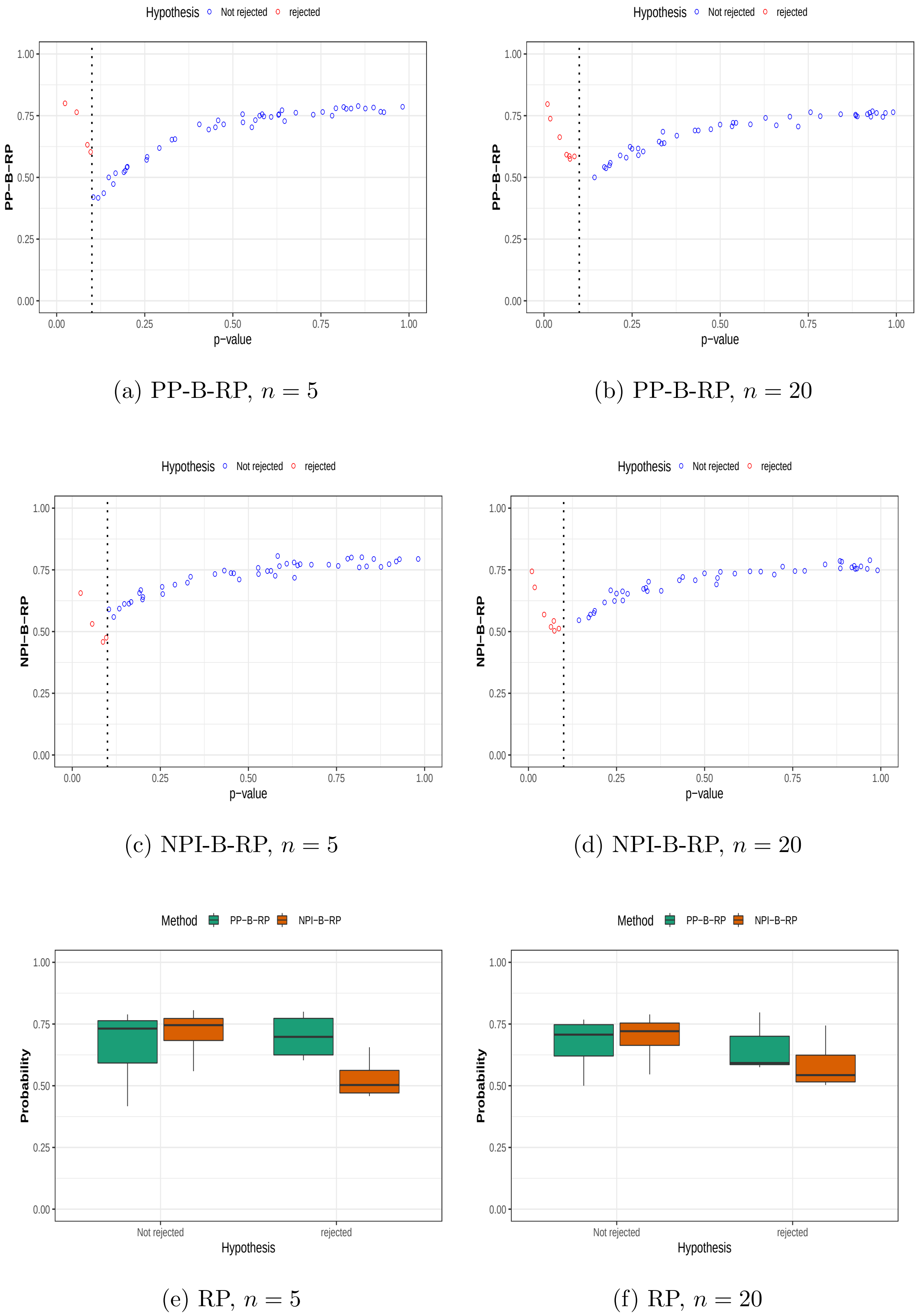
Figure 4.3: Simulations under H_0: values of PP-B-RP and NPI-B-RP for two-sample t-test, where $n = 5, 20$.

presented in Figure 4.3. The boxplots of RP values are shown for the two methods in both cases of rejection and non-rejection.

The values of RP for both methods tend to increase as the test statistic moves away from the test thresholds, regardless of the decision on H_0. It is expected and rational, as we discussed previously in Section 4.3. Increasing the size of samples leads to PP-B-RP and NPI-B-RP becoming close to 0.5 in both cases of rejection and non-rejection when the observed test statistics are close to the test thresholds. Also, the values of NPI-B-RP fluctuate narrowly as the sample size increases. These results happen with increasing the size of samples due to the variability of the bootstrap samples decreasing and the power of the test increasing. Simulation studies show that values of PP-B-RP have less variability than NPI-B-RP values, in particular when the sample size is small, as a result of the parametric model assumed for PP-B.

There is a tendency for PP-B-RP to be higher in cases of rejection than in non-rejection, whereas NPI-B-RP seems to be lower in cases of rejection than non-rejection. The reason for this is that the sample variance is included in the denominator of the test statistic for the two-sample t-test. The variance of PP-B is generally less than NPI-B due to the assumption of a parametric model in PP-B. For the upper tail test, PP-B samples lead to larger test statistic values than NPI-B samples, as a result of a smaller variance value in the denominator. Therefore, the PP-B-RP tends to be lower in non-rejection cases due to the computed test statistic from PP-B samples tending to lie in the rejection region. Conversely, PP-B-RP tends to be higher in the case of rejection than NPI-B-RP because we obtain more cases that reject H_0. It is similar to what was discussed in Section 4.3 for the upper tail one-sample t-test. We can observe a similar impact on patterns of RP values based on PP-B and NPI-B for the lower tail test. It is important to note that the lower tail two-sample t-test has negative values, which implies that PP-B samples lead to smaller test statistic values compared to NPI-B samples. Hence, we obtain a similar result to the upper tail test for PP-B-RP and NPI-B-RP.

Now, we consider the RP of the two-sample t-test when both samples are normally distributed with unequal variances. The procedure for determining the RP of Welch's t-test follows the same steps as for the two-sample t-test, except that we



(a) PP-B-RP, $n = 5$

(b) PP-B-RP, $n = 20$

(c) NPI-B-RP, $n = 5$

(d) NPI-B-RP, $n = 20$

(e) RP, $n = 5$

(f) RP, $n = 20$

Figure 4.4: Simulations under H_0: values of PP-B-RP and NPI-B-RP for Welch's t-test, where $n = 5, 20$.

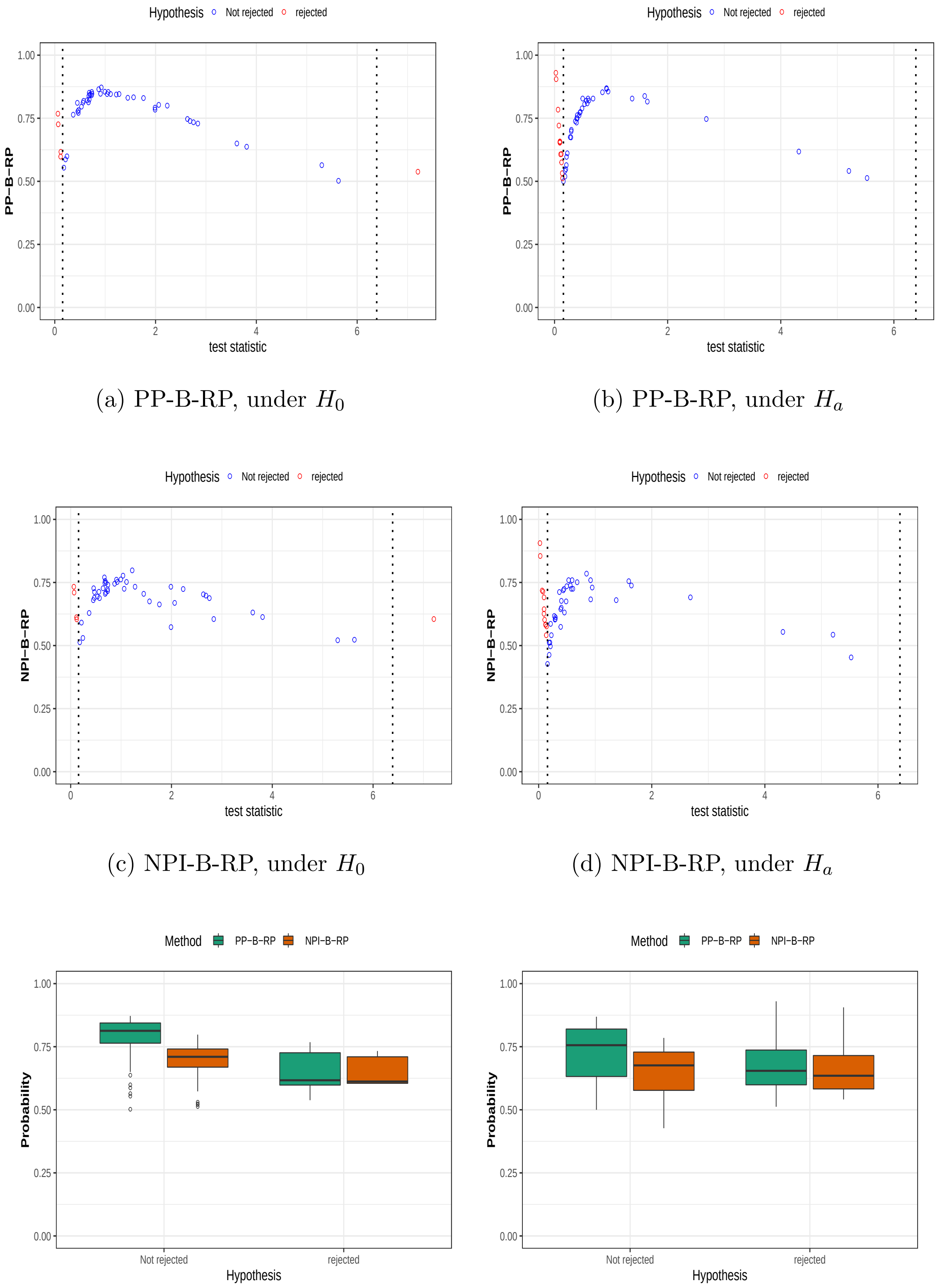
draw two original samples from Normal distributions with different standard deviations. Two samples of size n are simulated from two Normal distributions with different standard deviations, $\sigma_1 = 1$ and $\sigma_2 = 2$, but both with mean 2. A critical value of the test statistic for Welch's t-test is computed using the degrees of freedom which are random variables dependent on the size and variance of the sample. Therefore, we use the p-value for better visualization of figures rather than the critical value because each simulated sample has a different critical value even though all samples have the same size. The p-values and critical values are two different approaches that lead to the same result regarding whether the null hypothesis is rejected or not. Figure 4.4 shows the results of RP values for Welch's t-test using the two bootstrap methods with samples of size $n = 5, 20$ under H_0. The boxplots of RP values based on PP-B and NPI-B are displayed in both cases of rejection and non-rejection.

The values of RP for both methods tend to increase with increasing distance between the observed p-value and the test threshold, whatever the H_0 decision. We observe similar results as for the two samples with the Student's t-test presented before in this section. The parametric model assumed for PP-B results in lower variability of PP-B-RP values than NPI-B-RP values, especially when the sample size is small. The PP-B-RP seems to be greater in rejection cases than in non-rejection. In contrast, NPI-B-RP tends to be lower in the case of rejection compared to non-rejection. As the sample size increases, PP-B-RP and NPI-B-RP become closer to 0.5 in both cases of rejection and non-rejection when the observed p-value is close to the test threshold. The fluctuation in NPI-B-RP values is reduced with the increasing size of samples.

4.5 Bootstrap-RP for the F-test

In this section, we study the RP of the F-test using two bootstrap methods. The two-sample t-test requires random sampling from two normal populations with equal variances, while Welch's t-test is used in the case of unequal variances. The F-test is conducted to test the assumption of equal variances between two normal populations and, based on the results, a two-sample t-test or Welch's t-test can be used. A normal distribution of data is a prerequisite for conducting these parametric tests. The two sided F-test is considered, $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$, and level of significance $\alpha = 0.10$. Simulation studies are conducted to evaluate the performance of the two bootstrap methods for RP of the F-test by following the same steps as for the two-sample t-test in Section 4.4. We simulate two samples of size $n = 5$ under both H_0 and H_a in total $N = 50$ times. Under H_0 we generate data for the two original samples from the same Normal distribution with mean 0 and standard deviation 1. Under H_a we generate data from the two Normal distributions with different standard deviations, $\sigma_1 = 1$ and $\sigma_2 = 1.5$, but both with the same mean 0. For each of $N = 50$ samples, the observed test statistic and Bootstrap-RP were determined. It is important to note that the same data sets are used to compute the RP values for the F-test based on the two bootstrap methods, each with $B = 1000$ bootstrap samples. Also, the bootstrap samples for each method are the same size as the original sample. Figure 4.5 shows the results of RP values using PP-B and NPI-B methods under H_0 and H_a for samples of size $n = 5$. The boxplots of RP are presented for both rejections and non-rejections based on the two bootstrap methods.

The Bootstrap-RP values tend to be higher at the lower test threshold for both rejection and non-rejection cases as the impact of F-test follows F-distribution with small degrees of freedom. The simulations were performed by sampling under the alternative hypothesis due to more cases of test statistics close to the lower test threshold. This helps us to observe how the bootstrap methods perform for the RP of the F-test as test statistics become closer to the lower test threshold. The PP-B-RP becomes close to 0.5 in both cases of rejection and non-rejection when the observed test statistics are very close to the lower test threshold. The NPI-B-RP is substantially below 0.5 in some cases of non-rejection when test statistics are extremely close to the lower test threshold. The parametric model assumed for PP-B reduces the variability of RP values as shown in simulation studies. The RP value based on NPI-B fluctuates clearly because a parametric model is not assumed in this bootstrap method.



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

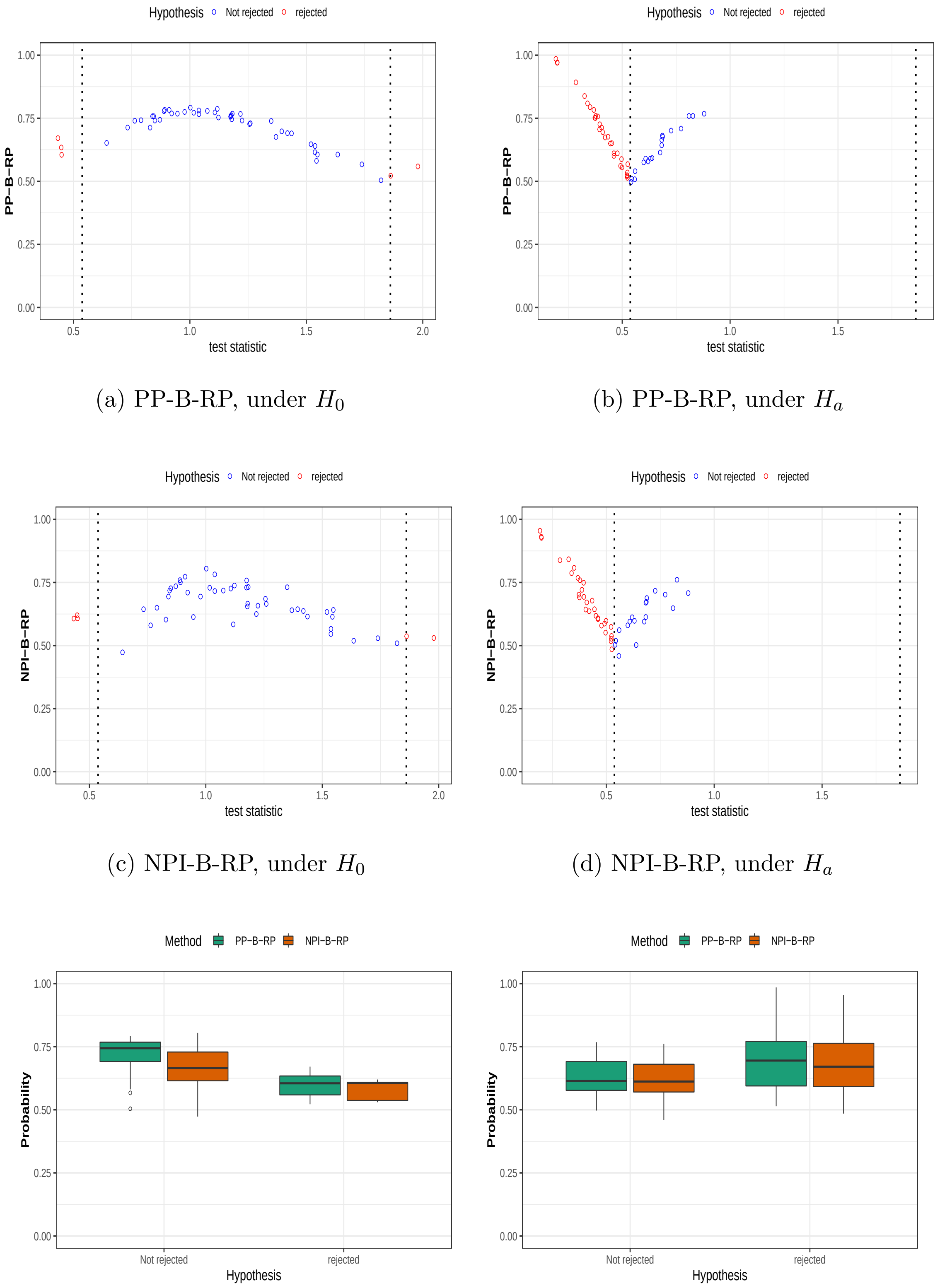
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_a

(e) RP, under H_0

(f) RP, under H_a

Figure 4.5: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for F-test, where $n = 5$.



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_a

(e) RP, under H_0

(f) RP, under H_a

Figure 4.6: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for F-test, where $n = 30$.

A larger sample size is considered in order to study the effect of increased sample size on Bootstrap-RP values for the F-ratio test. Figure 4.6 presents the results of RP values using the PP-B and NPI-B methods for samples of size $n = 30$ under H_0 and H_a. The boxplots of RP values for the two bootstrap methods are shown in both cases of rejection and non-rejection. As the size of samples increases, the pattern of RP values changes under both the null and alternative hypotheses. We observe a change in the pattern of the RP values obtained through simulations under H_0 as the impact of F-test follows F-distribution with larger degrees of freedom. Increasing the size of samples leads to increasing the power of the test, so we obtain more cases rejecting H_0 when simulations are performed by sampling under the alternative hypothesis. Simulations under H_a show changes in the pattern of the RP values as a result of changes in the observed test statistics in relation to the test threshold, as well as the effects of the F-test following the F-distribution with larger degrees of freedom. It is noteworthy that the variability of NPI-B-RP values is not reduced by increasing the size of samples. Additional simulation results for the NPI-B-RP of the F-test are provided in Appendix B, showing that the NPI-B-RP has high fluctuation even with increasing the size of samples. The test statistic for the F-test is computed from only the ratio of two sample variances. The NPI-B method has higher variability compared to the PP-B method because it does not use an assumed parametric model. Consequently, the fluctuation in NPI-B-RP values for the F-test does not decrease with the increasing size of samples due to the test statistic used is only the ratio of two sample variances.

4.6 NPI-RP and Bootstrap-RP for the likelihood ratio test

In this section, we study the RP of the likelihood ratio tests using the bootstrap method to compare with NPI-RP. The reproducibility probability of a test based on the NPI approach (NPI-RP) considers the test result for a predicted future sample of the same size as the original sample, this method is described in detail in Section 2.3. The exact NPI lower and upper reproducibility can only be computed for small data sets. Coolen and Marques [17] propose an alternative computational method to approximate NPI-RP for larger sample sizes via sampling of future orderings instead of considering all different possible orderings. They introduced sampling of orderings for the likelihood ratio test in order to overcome computational difficulties. In our work, we do not compute lower and upper reproducibility probabilities for the tests because it is hard to derive the minimum and maximum values of some test statistics, such as the test statistic of t-test, which depend on both the sample mean and variance. However, we can construct the confidence interval for the single value of Bootstrap-RP using formula $\hat{p} \pm z^{(1-\alpha/2)}\sqrt{\hat{p}(1-\hat{p})/n}$, where the proportion $\hat{p}$ is the predictied Bootstrap-RP value. Here we investigate whether or not the Bootstrap-RP tends to provide a value within the lower and upper NPI-RP.

Coolen and Marques [17] introduced sampling of future orderings for likelihood ratio tests with the test criterion in terms of the sample mean. The likelihood ratio test in the following test criterion involves the mean of the observed values. The null hypothesis H_0 is considered with a one sided alternative hypothesis, $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$, leading to the test criterion, H_0 being rejected if and only if

$$\frac{1}{n}\sum_{i=1}^{n} x_i > c \tag{4.5}$$

where c dependent on the significance level of the test and the assumed statistical model.

We cannot derive a precise value for the mean of a specific ordering O_j of the n future observations in the NPI approach because we do not assume precise values within the intervals $(x_{(i-1)}, x_{(i)})$. Therefore, the maximum lower bound and minimum upper bound for the mean corresponding to O_j can only be derived, which are denoted by $\underline{m}_j$ and $\overline{m}_j$, respectively. These are derived as follows

$$\underline{m}_j = \frac{1}{n}\sum_{i=1}^{n+1} s_i^j x_{(i-1)} \tag{4.6}$$

$$\overline{m}_j = \frac{1}{n}\sum_{i=1}^{n+1} s_i^j x_{(i)} \tag{4.7}$$

Suppose that the original data sample of size n led to rejection of H_0, so its mean exceeds c. In this case, the test result is reproduced if the future sample also rejects H_0. This occurs certainly for ordering O_j if $\underline{m}_j > c$, while it certainly does not occur if $\overline{m}_j \leq c$. However, we are unable to decide whether or not the original test result is reproduced if $\underline{m}_j \leq c < \overline{m}_j$. The NPI lower and upper probabilities for test reproducibility is derive for the case that the original data reject H_0 as

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j > c\} \tag{4.8}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j > c\} \tag{4.9}$$

where $j = 1, \ldots, \binom{n+m}{n}$ and $\mathbf{1}\{A\}$ is the indicator function which is equal to 1 if A is true and 0 else.

The similar arguments are followed when the original data do not reject H_0 to derive NPI lower and upper probabilities for test reproducibility as

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j \leq c\} \tag{4.10}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j \leq c\} \tag{4.11}$$

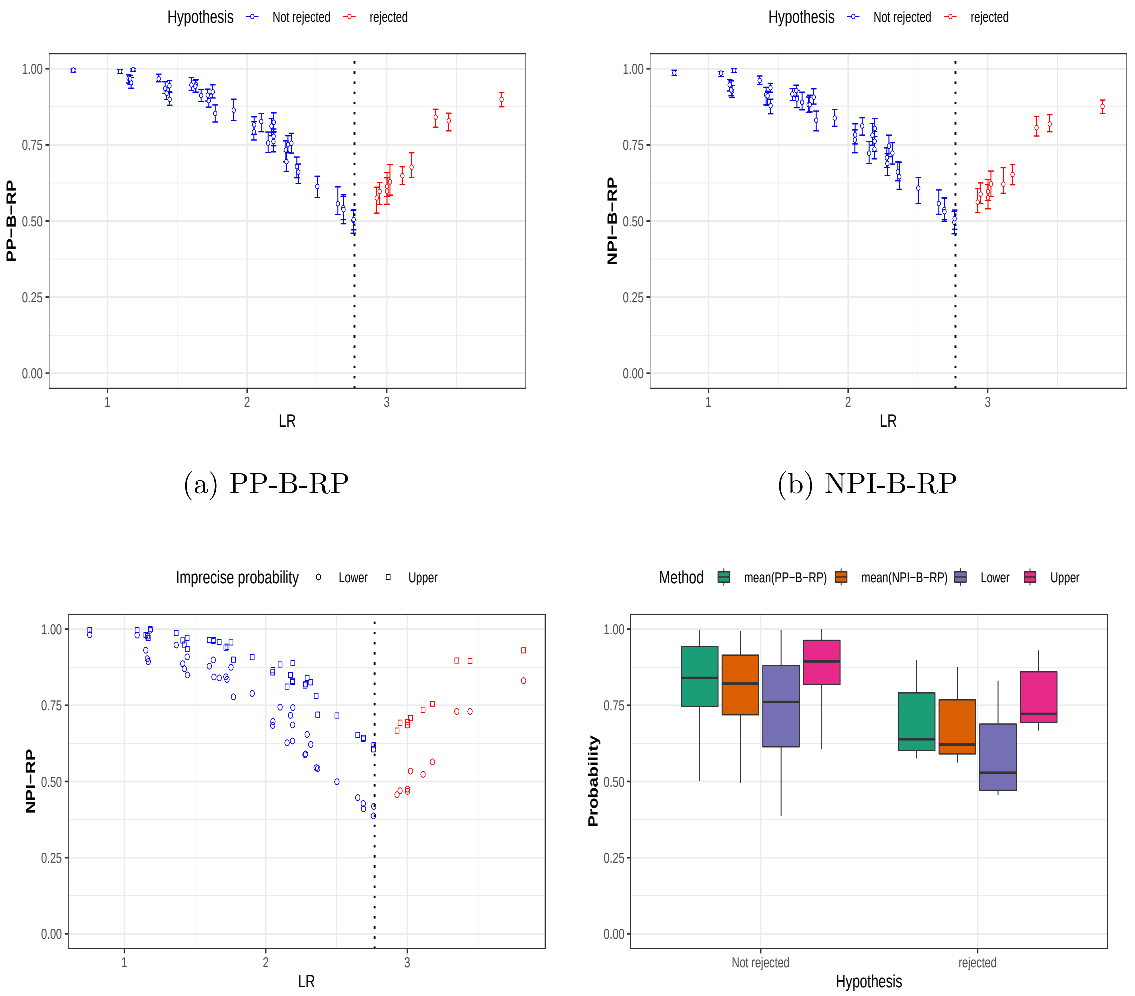
The decision rule may be expressed with the test criterion in terms of the sample mean $\overline{X}$ for the likelihood ratio test as test criterion (4.5), which rejects the null hypothesis for a significance level α if

$$\overline{X} > q_{(1-\alpha)} \tag{4.12}$$

where $q_{(1-\alpha)}$ is the $(1-\alpha)$ quantile of $\overline{X}$. It is well known that for independent and identically distributed $X_i \sim N(\mu, \sigma^2), i = 1, \ldots, n$, the distribution of the mean is

$$\overline{X} \sim N(\mu, \sigma^2/\sqrt{n})$$

We consider likelihood ratio tests for the mean value underlying the Normal population. For distributions with infinite range we have to define bounds of possible values for the future observations, which we denote by $x_{(0)} = L$ and $x_{(n+1)} = R$. It is obvious that we must assume values $L < x_{(1)}$ and $x_{(n)} < R$ such that the observations are within this range $[L, R]$, where L and R can depend on the actual



(a) PP-B-RP

(b) NPI-B-RP

(c) NPI-RP

(d) RP

Figure 4.7: Simulations under H_0: values of PP-B-RP, NPI-B-RP and NPI-RP for likelihood ratio test, where $n = 25$.

data observations. For n data observations $x_1 < x_2 < \ldots < x_n$, the lower and upper limits may be defined as $L = x_{(1)} - \frac{x_{(n)}-x_{(1)}}{n-1}$ and $R = x_{(n)} + \frac{x_{(n)}-x_{(1)}}{n-1}$.

We simulated $N = 50$ samples of size $n = 25$ from the Normal distribution with mean 2 and standard deviation 3 under H_0. We approximate NPI-RP for larger sample sizes via sampling of orderings instead of considering all different possible orderings. To achieve reasonable results, Coolen and Marques [17] suggest that the number of orderings sampled should be at least 2000. Considering the number of orderings sampled equal to 2000, the upper and lower RP for each of $N = 50$ samples were calculated based on the decision rule given in (4.12) with the level of significance

$\alpha = 0.10$. The NPI lower and upper reproducibility probabilities are calculated for rejection cases using Equations (4.8) and (4.9). In the case of non-rejection, we compute the NPI lower and upper reproducibility probabilities using Equations (4.10) and (4.11). We investigate whether or not the Bootstrap-RP methods tend to provide values that fall within the lower and upper NPI-RP for the likelihood ratio test. The RP for each of $N = 50$ samples was computed based on the PP-B and NPI-B methods using $B = 1000$ bootstrap samples. For each simulated sample, we compute RP values based on the bootstrap method and repeat the procedure 100 times, so we obtain $RP_1, \ldots, RP_{100}$. Then, we examine whether these values are between the corresponding lower and upper NPI-RP results. The same simulated samples are used to compute the RP values of the likelihood ratio test based on different bootstrap methods and NPI-RP. The observed likelihood ratio statistic, Bootstrap-RP, and NPI-RP were determined for each of $N = 50$ samples.

Figure 4.7 presents RP values using different bootstrap methods and NPI-RP under H_0 for samples of size $n = 25$. The minimum, mean and maximum values of 100 Bootstrap-RP for each simulated sample are computed. The boxplots of RP are displayed for both rejections and non-rejections based on the mean of PP-B-RP and NPI-B-RP, as well as the lower and upper NPI-RP. We found 90% of PP-B-RP values and 88% of NPI-B-RP values are included in the bounds of NPI-RP. We conclude that both PP-B-RP and NPI-B-RP results are consistent with NPI-RP because most of these values are located in the corresponding NPI-RP boundaries. The PP-B-RP and NPI-B-RP are in line with NPI-RP in terms of investigating test reproducibility as a prediction problem rather than an estimation problem. Further simulations were performed under H_a leading to similar results as for the case presented under H_0.

The two-sided for the likelihood ratio test, $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$, may be implemented in a similar procedure. The test criterion based on sample mean X reject the null hypothesis at a significant level if

$$\overline{X} < q_{(\alpha/2)} \quad \vee \quad \overline{X} > q_{(1-\alpha/2)} \tag{4.13}$$

where $q_{(\alpha/2)}$ and $q_{(1-\alpha/2)}$ are the $(\alpha/2)$ and $(1-\alpha/2)$ quantile of $\overline{X}$.

The minimum upper bound and maximum lower bound for the mean corresponding to O_j are remain unchanged as in Equations (4.6) and (4.7), respectively. In the case of a two-sided test, the NPI lower and upper probabilities are different because it needs to account for the two rejection regions. If the original data reject H_0, then the lower and upper RPs are derived as follows

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j > q_{(1-\alpha/2)} \quad \vee \quad \overline{m}_j < q_{(\alpha/2)}\} \tag{4.14}$$

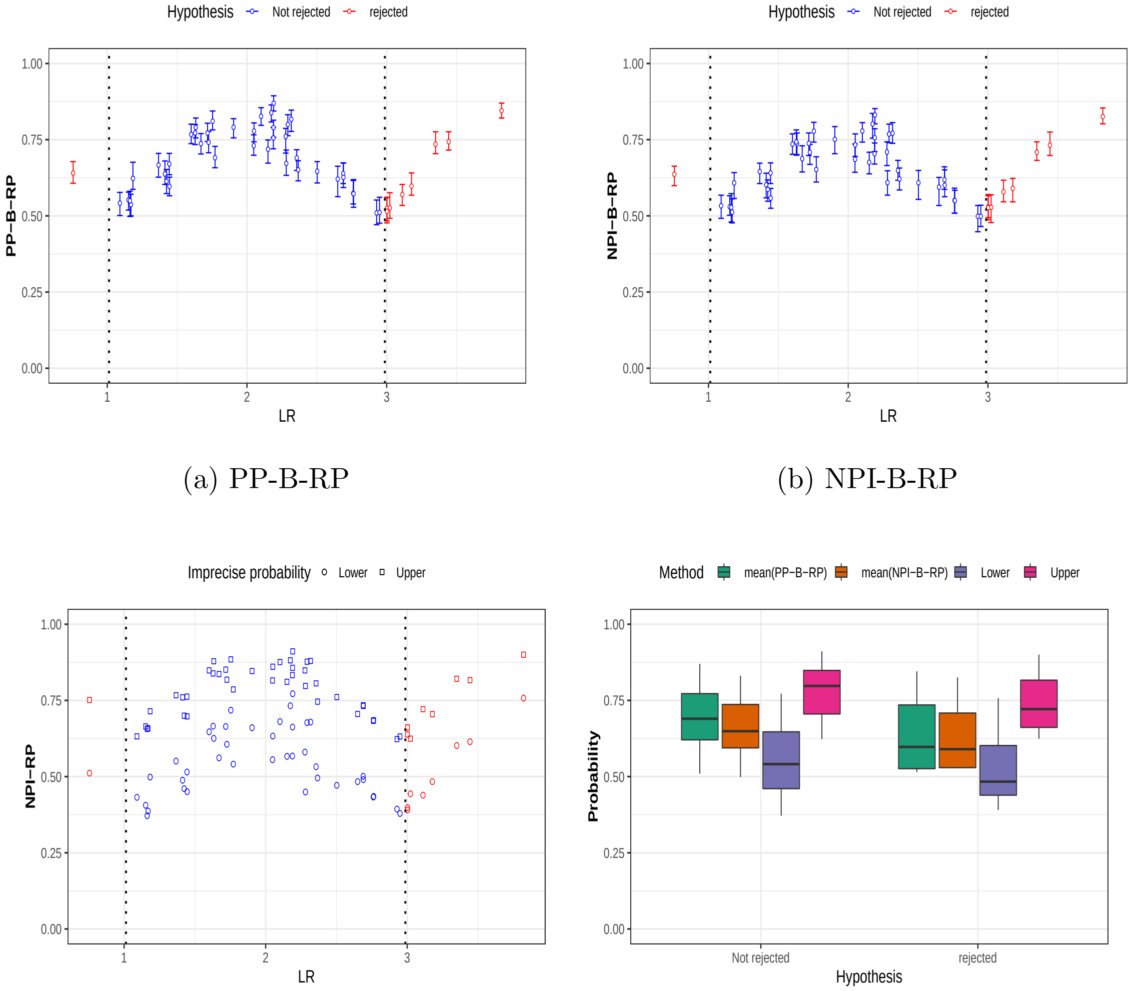
$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j > q_{(1-\alpha/2)} \quad \vee \quad \underline{m}_j < q_{(\alpha/2)}\} \tag{4.15}$$

If the decision for the original data is not rejected the null hypothesis, we have

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\underline{m}_j > q_{(\alpha/2)} \quad \wedge \quad \overline{m}_j < q_{(1-\alpha/2)}\} \tag{4.16}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_j \mathbf{1}\{\overline{m}_j > q_{(\alpha/2)} \quad \wedge \quad \underline{m}_j < q_{(1-\alpha/2)}\} \tag{4.17}$$

We have simulated $N = 50$ samples of size $n = 25$ from the Normal distribution with mean 2 and standard deviation 3 under H_0. For each case, we compute the lower and upper RPs for two sided test based on the decision rule given in (4.13) with the level of significance $\alpha = 0.10$ by considering the number of orderings sampled equal to 2000. The lower and upper reproducibility probabilities of the NPI are computed for rejection cases using Equations (4.14) and (4.15). In the case of non-rejection, we calculate the NPI lower and upper reproducibility probabilities based on Equations (4.16) and (4.17). The same simulated samples are used to compute the RP values based on the bootstrap and NPI methods. We compute RP values for the two sided test based on the bootstrap method and repeat the procedure 100 times for each simulated sample as we did with the one sided test. Figure 4.8 shows RP values for the likelihood ratio test with two sided alternative using different bootstrap methods and NPI-RP under H_0 for samples of size $n = 25$. For each simulated sample, the minimum, mean, and maximum Bootstrap-RP values are computed. The boxplots of RP are shown in both cases of rejection and non-rejection based on the mean of PP-B-RP and NPI-B-RP, along with the lower and upper NPI-RP. All values



(a) PP-B-RP

(b) NPI-B-RP

(c) NPI-RP

(d) RP

Figure 4.8: Simulations under H_0: values of PP-B-RP, NPI-B-RP, and NPI-RP for likelihood ratio test, where $n = 25$.

of PP-B-RP and NPI-B-RP are included in the bounds of NPI-RP, indicating that these bootstrap methods are in line with the reproducibility probability based on the NPI approach. Further simulations were performed under H_a yielding similar results to the case presented under H_0.

4.7 Concluding remarks

In this chapter, we present the PP-B method for the reproducibility of some parametric tests. We also provide a comparison through simulation studies with a similar predictive bootstrap method for test reproducibility, NPI-B. The test reproducibility is more naturally considered as a prediction problem than as an estimation problem. The explicit predictive nature of PP-B and NPI-B, which consider future observations, aligns well with the nature of test reproducibility. The reproducibility of tests has been studied using the PP-B and NPI-B methods via simulation studies. The RP values obtained with PP-B have less variability than those obtained with NPI-B, as a result of using an assumed parametric model for PP-B. Increasing sample size reduced the fluctuation of NPI-B-RP values because bootstrap samples became less variable and the power of test increased. However, the variability of NPI-B-RP values for the F-test is not reduced with increasing the size of samples because the test statistic for the F-test is calculated using only the ratio of two sample variances. We consider PP-B and NPI-B for the reproducibility of some parametric tests, but they can be applied to a wide range of parametric statistical tests.

The use of the bootstrap to predict RP avoids the hard calculations of the lower and upper boundaries in NPI-RP, as well as it is a flexible approach to use when considering large sample sizes. The Bootstrap-RP uses the point estimate to present the RP instead of the lower and upper values of NPI-RP, but we can construct the confidence interval for the single value of Bootstrap-RP. We explore whether or not the RP values using PP-B and NPI-B tend to be between the lower and upper NPI-RP for the likelihood ratio test. The predicted values of PP-B-RP and NPI-B-RP for the likelihood ratio test are mostly included within the bounds of NPI-RP, meaning these bootstrap methods are consistent with the NPI-RP approach. The PP-B-RP, NPI-B-RP, and NPI-RP consider test reproducibility from a predictive standpoint, which provides an appropriate formulation for inferring the RP of a test. It seems logical and natural to study the RP of a test with the same size of samples and significance level as in the actual test. Senn [64] discussed the circumstances in the real world may vary among different tests, including the size of samples. The bootstrap method for the reproducibility of tests can be extended to consider future sample sizes that are different from the data sample size or use varying levels of statistical significance. However, the use of the same sample sizes and significance levels as in the actual test is logical from the perspective of theoretical reproducibility, particularly within a frequentist statistical framework.

Chapter 5

Misspecification of Parametric Predictive Bootstrap

5.1 Introduction

The parametric predictive bootstrap method relies on assumed parametric models. This method samples all bootstrap observations based on an assumed distribution with estimated parameters from the available data. In real-world applications, data sets are never perfect and could suffer from several problems, one of these problems is model misspecification. We refer to the misspecified PP-B model for situations that occur when the distribution of data sets used is incorrectly specified. As a consequence, the behaviour of PP-B may be different from the one that we would expect on the basis of a well-specified model. In this chapter, three different scenarios are considered to examine the performance of PP-B regarding the assumed parametric model. In the first scenario, the PP-B samples are generated from the same distribution as the original data set, meaning the distribution assumed for the PP-B method is well-defined. This scenario is referred to as PP-B$_{(1)}$. In the second scenario, we generate PP-B samples from a different distribution than that used for the original data set, but they are closely related to one another. This scenario is referred to as PP-B$_{(2)}$. In the third scenario, the PP-B samples are generated from a completely different distribution than that used to sample the original data set. This scenario is referred to as PP-B$_{(3)}$.

This chapter is organized as follows: In Section 5.2, we investigate the impact of the misspecified PP-B model on the performance of the LC prediction interval. In Section 5.3, the performance of PP-B is compared with different bootstrap methods using confidence regions. Also, we study the effect of the misspecified PP-B model on the performance of the confidence regions. In Section 5.4, we illustrate how to apply Banks' comparison method for prediction intervals to compare PP-B with other bootstrap methods. We then examine the effect of a misspecified PP-B model the performance of the prediction regions. In Section 5.5, we study how the fixed bootstrap variance of PP-B impacts the reproducibility probability of one-sample t-test. In the final section, some concluding remarks are provided.

5.2 LC prediction interval

The aim of this section is to study the effect of the misspecified PP-B model on the performance of the LC prediction interval. In Chapter 3, we illustrated how to construct LC prediction intervals for the mean and variance of m future observations, as well as for a single future observation. Those intervals can be used to evaluate the prediction performance of different bootstrap methods. It has been shown that predictive bootstrapping methods, such as PB-B and NPI-B, perform well and provide a good proportion of coverage for LC prediction intervals. It is desirable to have a proportion of coverage that is close to nominal coverage probability, along with a shorter interval width. PP-B has the advantage of providing good coverage with a shorter average width of intervals than NPI-B.

We consider three different scenarios to examine PP-B's performance based on the assumed parametric model with LC prediction interval. These scenarios are examined for a variety of original sample sizes $n = 50, 100, 200, 400$ with confidence levels 95% and 90%. The original data sets are simulated from an Exponential distribution with rate parameter $\lambda = 5$. In the first scenario, the PP-B samples are generated using the same distribution as the original data set, which means the distribution assumed for the PP-B method is well-defined. In the second scenario, we generate PP-B samples from the Gamma distribution with two parameters α and

β. In statistics, there are several relationships among probability distributions can be categorized in various ways such as one distribution is a special case of another with a broader parameter space. The Exponential distribution is a special case of the Gamma distribution with shape $\alpha = 1$ and rate $\beta = \lambda$. In the third scenario, the PP-B samples are generated from the Normal distribution with parameters μ and σ^2. The Normal distribution differs greatly from the Exponential distribution, including the fact that it is symmetric about the mean, has a bell-shaped curve, and has a domain of all real numbers. The second and third scenarios illustrate the performance of the PP-B method when the model assumed in PP-B is different from the actual model assumed in the original data. An evaluation of each scenario is based on the coverage proportion and the average width of intervals. It is not necessarily better to have narrower intervals, but the width is only important if the coverage proportion are accurate.

The simulation study is carried out to investigate the proportion of coverage and the average interval widths under the three scenarios. The PP-B method involves estimating the parameters of an assumed distribution using available data. In each scenario, we estimate parameters based on the model assumed for PP-B. The LC prediction intervals are constructed for the mean of $m = n, n/2$ future observations. A past sample of size n and a future sample of size m are generated independently from Exp(5), then we compute the mean of the m future observations. In each scenario, we draw $B = 1000$ bootstrap samples of size m from the past sample and compute the mean of each bootstrap sample. After that, we use Equation (2.32) to construct the LC prediction interval for the mean of m future observations. Then, we determine if this interval contains the observed mean of the m future observations. This procedure is repeated $N = 1000$ times in order to see the coverage proportion and the average interval widths for each scenario. We also study the LC prediction intervals for a single future observation under three scenarios in terms of the coverage proportion and average width of intervals. The prediction interval for a single future observation is constructed as discussed in Section 2.6. Table 5.1 presents the coverage proportions and average interval widths for three different scenarios.

(a) $m = n$

Scenario	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B$_{(1)}$	CP_{LC}	0.9380	0.9630	0.9480	0.9480	0.8780	0.9160	0.8950	0.9050
	AW_{LC}	0.1553	0.1101	0.0778	0.0552	0.1294	0.0922	0.0653	0.0464
PP-B$_{(2)}$	CP_{LC}	0.9260	0.9540	0.9410	0.9400	0.8590	0.9060	0.8840	0.8960
	AW_{LC}	0.1520	0.1091	0.0777	0.0549	0.1262	0.0910	0.0650	0.0461
PP-B$_{(3)}$	CP_{LC}	0.9280	0.9610	0.9420	0.9400	0.8710	0.9130	0.8810	0.8970
	AW_{LC}	0.1533	0.1094	0.0776	0.0550	0.1282	0.0917	0.0652	0.0462

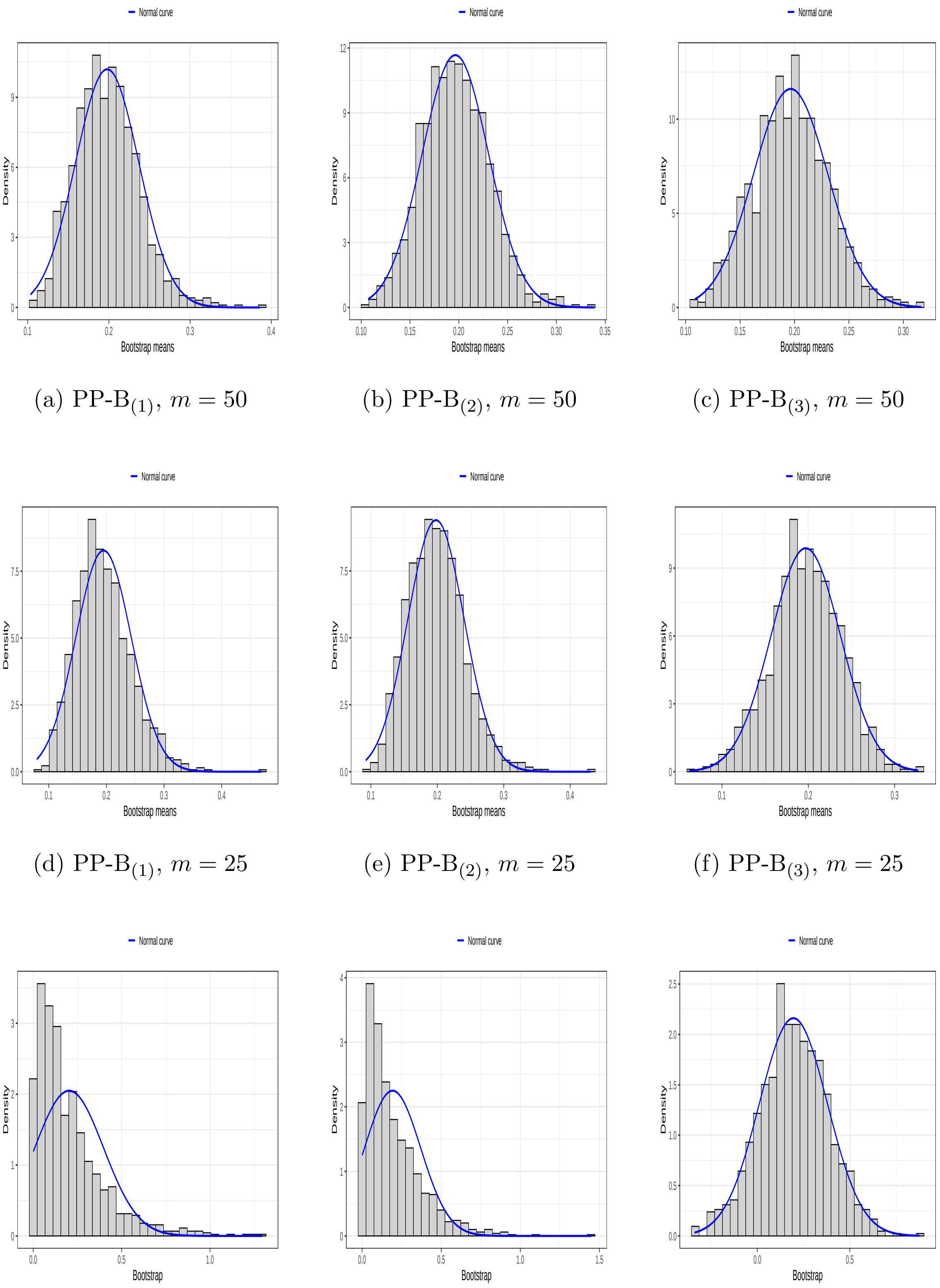
(b) $m = n/2$

Scenario	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B$_{(1)}$	CP_{LC}	0.9370	0.9340	0.9580	0.9460	0.8800	0.8740	0.9070	0.8980
	AW_{LC}	0.1898	0.1348	0.0952	0.0675	0.1582	0.1127	0.0799	0.0567
PP-B$_{(2)}$	CP_{LC}	0.9290	0.9270	0.9550	0.9490	0.8690	0.8670	0.9050	0.9010
	AW_{LC}	0.1864	0.1333	0.0950	0.0673	0.1546	0.1113	0.0795	0.0565
PP-B$_{(3)}$	CP_{LC}	0.9230	0.9410	0.9580	0.9470	0.8800	0.8740	0.9100	0.9010
	AW_{LC}	0.1875	0.1340	0.0952	0.0673	0.1568	0.1123	0.0798	0.0566

(c) $m = 1$

Scenario	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B$_{(1)}$	CP_{LC}	0.9340	0.9340	0.9360	0.9410	0.8920	0.8860	0.8920	0.8900
	AW_{LC}	0.7312	0.7305	0.7279	0.7290	0.5900	0.5893	0.5874	0.5882
PP-B$_{(2)}$	CP_{LC}	0.9260	0.9280	0.9370	0.9390	0.8760	0.8740	0.8900	0.8860
	AW_{LC}	0.7200	0.7228	0.7259	0.7274	0.5804	0.5824	0.5840	0.5863
PP-B$_{(3)}$	CP_{LC}	0.9290	0.9250	0.9330	0.9360	0.9060	0.9080	0.9130	0.9150
	AW_{LC}	0.7711	0.7761	0.7784	0.7794	0.6483	0.6524	0.6544	0.6552

Table 5.1: *Coverage of $100(1-2\alpha)\%$ prediction interval for the mean of $m = n, n/2$ future observations and for a single future observatio $m = 1$ under the three scenarios.*



(a) PP-B$_{(1)}$, $m = 50$

(b) PP-B$_{(2)}$, $m = 50$

(c) PP-B$_{(3)}$, $m = 50$

(d) PP-B$_{(1)}$, $m = 25$

(e) PP-B$_{(2)}$, $m = 25$

(f) PP-B$_{(3)}$, $m = 25$

(g) PP-B$_{(1)}$, $m = 1$

(h) PP-B$_{(2)}$, $m = 1$

(i) PP-B$_{(3)}$, $m = 1$

Figure 5.1: The bootstrap histograms of different scenarios for the future sample mean and single future observation, where $n = 50$ and $m = 50, 25, 1$.

It is apparent that all different scenarios provide coverage that is close to the nominal coverage probability with similar average interval widths. These results occur for future sample means due to the central limit theorem (CLT), which states that the sampling distribution of the sample mean becomes closer to a Normal distribution as the sample size gets larger, regardless of the distribution from which we are sampling [49]. The CLT ensures a good approximation with a sample size of 30, meaning that the distribution of the sample mean is close to the Normal distribution. The single future observation has the largest average interval widths compared to the future sample mean with different sizes $m = n, n/2$ due to the fact that the larger samples produce narrower intervals. As a result, all different scenarios have good coverage for a single future observation. Figure 5.1 shows a histogram of $B = 1000$ bootstrap replications of the future sample means and single future observation obtained by the three scenarios, with $n = 50$ and $m = 50, 25, 1$. The normal density curve is superimposed on each histogram. As the sample size grows, the histogram of the future sample means leads to graphed results that are closer to the normal density curve. In the case of a single future observation, the shape of the bootstrap histogram depends on the assumed distribution for each scenario from which PP-B is sampled.

Now, we consider the variance of $m = n, n/2$ future observations to evaluate the PP-B method under three scenarios. Simulations are conducted with the variance by applying the same scenarios, past samples, and confidence levels that were used in the future sample mean studies. Table 5.2 shows the results of coverage proportion and average width of intervals for the future sample variance based on the three scenarios. The first and second scenarios lead to good coverage proportions with similar average interval widths for all future sample sizes and confidence levels. In contrast, the third scenario performed poorly with under-coverage far from the nominal coverage probabilities for all cases considered. Figure 5.2 displays a histogram of $B = 1000$ bootstrap replications of the future sample variance for these three scenarios, with $n = 50$ and $m = 50, 25$. Histograms for the first and second scenarios are markedly non-normal, especially for smaller samples when $m = 25$, with a long tail toward the right. The shape of the bootstrap histogram for future

(a) $m = n$

Scenario	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B$_{(1)}$	CP_{LC}	0.9370	0.9560	0.9430	0.9470	0.8820	0.9040	0.8950	0.9040
	AW_{LC}	0.0759	0.0538	0.0380	0.0270	0.0612	0.0442	0.0316	0.0226
PP-B$_{(2)}$	CP_{LC}	0.9140	0.9330	0.9360	0.9490	0.8370	0.8770	0.8790	0.8870
	AW_{LC}	0.0835	0.0610	0.0437	0.0310	0.0649	0.0485	0.0355	0.0255
PP-B$_{(3)}$	CP_{LC}	0.6650	0.6840	0.6790	0.6760	0.5830	0.5790	0.6030	0.5810
	AW_{LC}	0.0437	0.0310	0.0220	0.0156	0.0362	0.0258	0.0184	0.0131

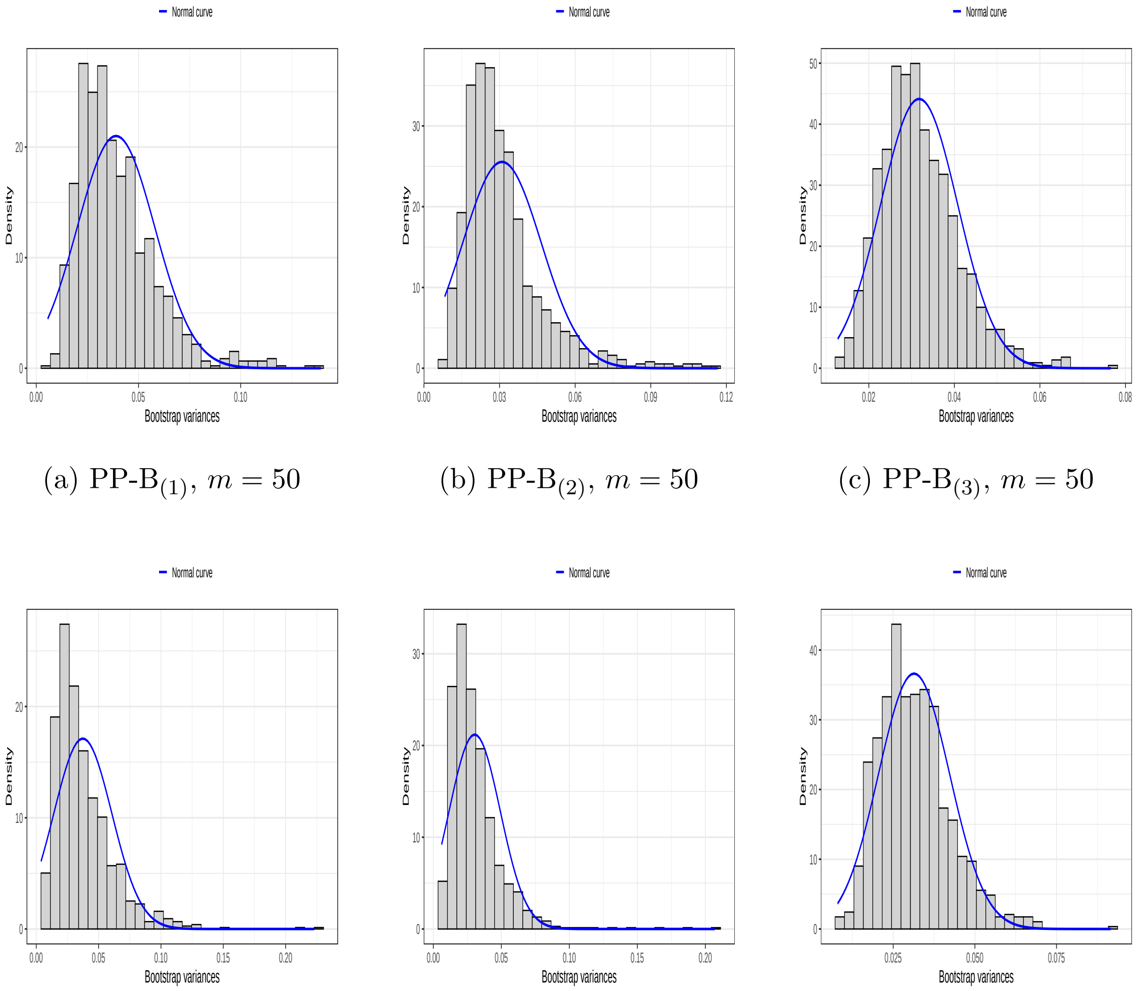
(b) $m = n/2$

Scenario	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B$_{(1)}$	CP_{LC}	0.9440	0.9420	0.9500	0.9420	0.8810	0.8870	0.9130	0.8840
	AW_{LC}	0.0953	0.0686	0.0486	0.0348	0.0757	0.0557	0.0401	0.0289
PP-B$_{(2)}$	CP_{LC}	0.9160	0.9260	0.9420	0.9380	0.8570	0.8590	0.8890	0.8780
	AW_{LC}	0.1008	0.0733	0.0531	0.0379	0.0777	0.0582	0.0430	0.0311
PP-B$_{(3)}$	CP_{LC}	0.6960	0.6730	0.6800	0.6700	0.5950	0.5950	0.5970	0.5980
	AW_{LC}	0.0537	0.0381	0.0269	0.0191	0.0443	0.0317	0.0225	0.0160

Table 5.2: *Coverage of* $100(1-2\alpha)\%$ *prediction interval for the variance of* $m = n, n/2$ *future observations under the three scenarios.*

sample variance depends on the assumed distribution for each scenario from which PP-B is sampled.

According to the central limit theorem, the sampling distribution of the sample mean approximately follows a normal distribution for large sample sizes. Therefore, the inference procedures for the mean were robust to violations of the normality assumption, in particular for large samples. Inference procedures for variance based on the assumption of a Normal distribution can perform very badly when this as-



(a) PP-B$_{(1)}$, $m = 50$

(b) PP-B$_{(2)}$, $m = 50$

(c) PP-B$_{(3)}$, $m = 50$

(d) PP-B$_{(1)}$, $m = 25$

(e) PP-B$_{(2)}$, $m = 25$

(f) PP-B$_{(3)}$, $m = 25$

Figure 5.2: The bootstrap histograms of different scenarios for the future sample variance, where $n = 50$ and $m = 50, 25$.

sumption is violated, even with large sample sizes. The original data sets come from an Exponential distribution, but for the PP-B method in the third scenario, we assume incorrectly that these data come from a Normal distribution. Therefore, we obtain coverage proportions that are far from nominal coverage probabilities for all cases of variance. In the second scenario, we generate PP-B samples from a different distribution of the original data sets, but they are closely related. Consequently, we achieve good coverage proportions, but the performance in the first scenario is mostly better for variance when the distribution assumed for the PP-B method is well-defined.

5.3 Confidence regions

In this section, we illustrate how to apply Banks' comparison method to evaluate the estimation performance of the different bootstrap methods using the confidence regions technique. We then examine the impact of the misspecified PP-B model on the performance of the confidence region. The main requirement for confidence regions is that the true coverage probability is close to the nominal coverage probability. This has motivated several simulations to focus on accuracy at custom sizes, such as 0.99, 0.95 and 0.90. Banks [4] investigated the global measure of coverage accuracy to compare different bootstrap methods. A total of 20 confidence regions are created, each with a nominal coverage probability of 0.05 by

$$CRL_{(i)} = \left(q_{(\frac{\alpha_{i+1}}{2})}, q_{(\frac{\alpha_i}{2})}\right) \tag{5.1}$$

$$CRR_{(i)} = \left(q_{(1-\frac{\alpha_i}{2})}, q_{(1-\frac{\alpha_{i+1}}{2})}\right) \tag{5.2}$$

where $i = 1, 2, \ldots, 10$, $\alpha_{i+1} = \alpha_i - 0.10$, $\alpha_1 = 1$ and $q_{(z)}$ is the z^{th} quantile of statistical values, so $CRL_{(i)}$ and $CRR_{(i)}$ are the confidence regions presenting the left tail and right tail of the global measure of coverage accuracy, respectively. The 10 confidence regions with a nominal coverage probability of 0.10 can be obtained using Equations (5.1) and (5.2) as follows:

$$CR_{(i)} = CRL_{(i)} \cup CRR_{(i)} \tag{5.3}$$

Both divisions of confidence regions are used to show the best bootstrap method that have the closest true coverage probability to the nominal coverage probability for a specific parameter of interest. Banks [4] used a chi-squared test of goodness of fit to assess the discrepancy in coverage proportion with different parameters, distributions and sample sizes, to compare his bootstrap method to other bootstrap techniques, e.g. Efron's method [30], Rubin's Bayesian bootstrap [63] and smoothed Rubin's bootstrap [4]. He considered the best bootstrap method to be the one having the lowest chi-squared (χ^2) values. We here intend to use the confidence regions technique for comparison of PP-B with other methods of bootstrap, described in Section 2.2. Also, we study the effects of the misspecified PP-B model on the performance of the confidence region.

The coverage proportions are estimated at 10 and 20 confidence regions of the bootstrap confidence interval for the mean with a sample size of 50 from Beta(3,1). A total of $N = 1000$ data sets are generated from Beta(3,1) with sample size $n = 50$. The different bootstrap methods are applied to each data set $B = 1000$ times. The means of the bootstrap samples are computed, and then we compute the 10 and 20 confidence regions using Equations (5.1), (5.2) and (5.3). Thereafter, we determine which confidence regions contain the true mean of the Beta(3,1) distribution. By repeating this procedure across all $N = 1000$ generated data sets, we are able to find the coverage proportions for the true mean in the 10 and 20 confidence regions. In Tables 5.3 and 5.4, we present the coverage proportions for the true mean in the 10 and 20 confidence regions, respectively.

The PP-B and NPI-B methods lead to coverage proportions far from the nominal level of 0.10 in most of the 10 confidence regions, and far from 0.05 in most of the 20 confidence regions. They produce wider confidence regions than other bootstrap methods due to the greater variability in their bootstrap samples. Therefore, the PP-B and NPI-B methods have over-coverage results for many confidence regions. In contrast, the PB and EB methods illustrate their superiority in making coverage proportions in each of the 10 and 20 confidence regions close to 0.10 and 0.05, respectively. The chi-square test can be used to assess the discrepancy between the nominal coverage probabilities and coverage proportions at distinct confidence levels based on different bootstrap methods. The first row of Table 5.5 shows the chi-squared values for the different bootstrap methods. The PB and EB methods are clearly superior at achieving the lowest discrepancy between nominal coverage probabilities and coverage proportions in both divisions of the confidence regions. Simulations were repeated several times and the results were consistent, as shown in Table 5.5. The PP-B method achieves lower chi-squared values than NPI-B for all cases, but the PB and EB methods are performing better.

We consider a variety of sample sizes to explore whether or not sample size affects the performance of different bootstrap techniques. Table 5.6 outlines the chi-squared values obtained from the coverage proportions for the mean at various sample sizes n based on different bootstrap methods. The chi-squared values for all

$CR_{(i)}$	PP-B	NPI-B	PB	EB
1	0.136	0.153	0.103	0.107
2	0.136	0.148	0.097	0.087
3	0.142	0.139	0.104	0.110
4	0.135	0.149	0.113	0.098
5	0.134	0.121	0.107	0.113
6	0.098	0.110	0.104	0.110
7	0.094	0.078	0.102	0.104
8	0.059	0.054	0.091	0.094
9	0.037	0.028	0.099	0.088
10	0.029	0.020	0.080	0.089

Table 5.3: *The coverage proportions for the mean in the 10 confidence regions, where* $n = 50$.

bootstrap methods show consistent results for all cases of different sample sizes in both confidence region divisions. As the sample size increases, the corresponding chi-squared values of the NPI-B method decrease in both divisions of the global measure of coverage accuracy. However, it still has a higher chi-squared value in all cases of confidence region divisions compared to the other bootstrap methods. In both the PP-B and the NPI-B methods, the chi-squared values are large because of the large discrepancies between the nominal coverage probabilities and coverage proportions. It appears that both the PB and EB methods achieve good coverage accuracy since their chi-squared values are low.

Additionally, we evaluate different bootstrap methods using confidence regions of the bootstrap confidence interval for the variance. The chi-squared goodness of fit values are computed based on the coverage proportions of the four bootstrap methods for variance at different sample sizes n. Several sample sizes are considered, to investigate whether there is an influence of sample size on the performance of different bootstrap methods or not. Table 5.7 shows the chi-squared values obtained

method	PP-B		NPI-B		PB		EB	
i	$CRL_{(i)}$	$CRR_{(i)}$	$CRL_{(i)}$	$CRR_{(i)}$	$CRL_{(i)}$	$CRR_{(i)}$	$CRL_{(i)}$	$CRR_{(i)}$
1	0.073	0.063	0.072	0.081	0.055	0.048	0.062	0.045
2	0.072	0.064	0.078	0.070	0.055	0.042	0.049	0.038
3	0.074	0.068	0.073	0.066	0.055	0.049	0.050	0.060
4	0.072	0.063	0.070	0.079	0.054	0.059	0.057	0.041
5	0.074	0.060	0.054	0.067	0.063	0.044	0.058	0.055
6	0.057	0.041	0.055	0.055	0.049	0.055	0.056	0.054
7	0.053	0.041	0.034	0.044	0.059	0.043	0.054	0.050
8	0.034	0.025	0.020	0.034	0.045	0.046	0.047	0.047
9	0.025	0.012	0.012	0.016	0.054	0.045	0.050	0.038
10	0.017	0.012	0.008	0.012	0.048	0.032	0.054	0.035

Table 5.4: *The coverage proportions for the mean in the 20 confidence regions, where* $n = 50$.

Repetition	10 CR				20 CR			
	PP-B	NPI-B	PB	EB	PP-B	NPI-B	PB	EB
1	171.90	233.44	6.10	6.06	184.76	245.64	21.16	14.16
2	174.92	229.76	12.36	8.08	188.00	238.60	23.00	16.92
3	170.06	235.52	5.28	9.90	183.72	247.80	13.92	23.24
4	174.36	233.94	10.44	7.46	189.76	243.60	25.88	18.36
5	180.8	229.06	9.28	8.00	192.76	236.92	16.36	19.64
6	172.84	234.66	7.66	5.36	184.24	243.60	17.92	16.16
7	169.10	229.40	6.34	7.08	183.32	236.76	16.24	17.20
8	174.20	234.88	9.20	4.28	187.76	249.60	18.12	15.84
9	174.30	228.46	11.58	10.28	188.60	240.60	20.08	19.16
10	174.74	235.70	9.56	4.86	186.16	253.24	19.16	12.60

Table 5.5: *The chi-squared values obtained from coverage proportions for the mean, where* $n = 50$.

n	measures	10 CR				20 CR			
		PP-B	NPI-B	PB	EB	PP-B	NPI-B	PB	EB
30	χ^2	161.42	230.56	14.80	16.98	187.08	249.648	27.72	29.76
	p-value	0.000	0.000	0.097	0.049	0.000	0.000	0.089	0.055
100	χ^2	116.40	154.10	6.26	6.78	132.52	168.88	25.44	21.64
	p-value	0.538	0.000	0.714	0.660	0.000	0.000	0.147	0.303
200	χ^2	137.78	145.28	3.16	2.18	156.04	157.20	16.84	14.60
	p-value	0.000	0.000	0.958	0.988	0.000	0.000	0.601	0.748
300	χ^2	115.92	129.14	10.24	11.00	125.84	142.16	22.60	25.72
	p-value	0.000	0.000	0.331	0.276	0.000	0.000	0.255	0.138

Table 5.6: *The chi-squared values for the mean and their p-values with different sample sizes n.*

from the coverage proportions for the true variance based on the different bootstrap procedures. As a result of the high discrepancy between the nominal coverage probabilities and coverage proportions, the χ^2 values for both the PP-B and the NPI-B methods are large. The PB method distributes the coverage proportions well over the 10 and 20 confidence regions, which is reflected in the low chi-squared value. This method performs better than other bootstrap methods in terms of reducing the discrepancy between nominal coverage probabilities and coverage proportions. The EB method leads to a high chi-squared value in the 20 confidence regions due to the large discrepancies between the nominal coverage probabilities and coverage proportions. For the variance, EB exhibits a noticeable decrease in its ability to distribute coverage proportions close to the nominal level of 0.05 in most of the 20 confidence regions. However, the corresponding chi-squared values of EB decrease as the sample size increases in the 20 confidence regions. It has been observed that when the sample size increases, the pattern of chi-squared values corresponding to NPI-B decreases in both confidence region divisions for the mean and variance. However, the chi-squared values of NPI-B are the largest in both divisions of the

n	measures	10 CR				20 CR			
		PP-B	NPI-B	PB	EB	PP-B	NPI-B	PB	EB
50	χ^2	133.62	302.88	11.26	29.14	200.24	349.04	49.56	107.44
	p-value	0.000	0.000	0.258	0.001	0.000	0.000	0.000	0.055
100	χ^2	119.14	199.36	12.46	20.02	158.48	245.32	54.64	86.40
	p-value	0.538	0.000	0.189	0.018	0.000	0.000	0.000	0.000
200	χ^2	156.16	190.90	3.06	7.42	175.24	240.72	33.68	37.28
	p-value	0.000	0.000	0.962	0.593	0.000	0.000	0.020	0.007
300	χ^2	122.62	155.38	14.80	16.06	147.20	183.12	26.24	35.88
	p-value	0.000	0.000	0.097	0.066	0.000	0.000	0.124	0.011

Table 5.7: *The chi-squared values for the variance and their p-values with different sample sizes n.*

confidence regions, followed by the PP-B method.

Now, we study the impact of the misspecified PP-B model on confidence regions' performance. Three scenarios are considered to examine PP-B's performance based on the assumed parametric model. These scenarios are investigated for several sample sizes $n = 50, 100, 200, 400$ from an Exponential distribution with rate parameter $\lambda = 5$. For each scenario, we compute the chi-squared goodness of fit values obtained from the coverage proportions for the mean and variance. We specify the parameters of the assumed distribution for each scenario in order to find the coverage proportion for the mean and variance. It is important to emphasize that the PP-B method generates the bootstrap sample by estimating the distribution parameters from the available data. In the first scenario, we assume that these data come from the Exponential distribution with rate parameter $\lambda = 5$, meaning that the distribution assumed for the PP-B method is well-defined. In the second scenario, we assume incorrectly that these data follow the Gamma distribution with parameters $\alpha = 1$ and $\beta = 5$. In the third scenario, these data are incorrectly assumed to come from the Normal distribution with parameters $\mu = 1/5$ and $\sigma^2 = 1/25$. It is important to

(a) mean

n	measures	10 CR			20 CR		
		PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$	PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$
50	χ^2	139.38	99.08	105.28	168.80	123.04	147.56
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
100	χ^2	179.16	150.00	157.00	194.16	167.80	178.44
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
200	χ^2	163.78	154.44	158.00	177.04	169.32	177.64
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
300	χ^2	147.56	137.80	143.68	156.80	150.16	164.88
	p-value	0.000	0.000	0.000	0.000	0.000	0.000

(b) variance

n	measures	10 CR			20 CR		
		PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$	PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$
50	χ^2	203.44	17.20	254.34	332.68	294.4	459.72
	p-value	0.000	0.046	0.000	0.000	0.000	0.000
100	χ^2	281.98	76.52	208.3	376.96	248.20	318.24
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
200	χ^2	313.80	109.12	197.20	370.24	222.00	244.00
	p-value	0.000	0.000	0.000	0.000	0.000	0.000
300	χ^2	309.94	102.50	252.08	340.36	181.20	292.92
	p-value	0.000	0.000	0.000	0.000	0.000	0.000

Table 5.8: *The chi-squared values for the mean and variance with their p-values under the three scenarios.*

note that the parameters are estimated based on the distribution assumed for PP-B in each scenario.

Table 5.8 presents the chi-squared values obtained from the coverage proportions for the true mean and variance based on the three scenarios. It is apparent that the chi-squared values of all different scenarios are large with only one exception for variance in the second scenario when $n = 50$ and the confidence level is divided into 10 confidence regions. It is not unexpected that PP-B does not lead to coverage proportions close to nominal levels of most confidence region divisions, as it is not developed for estimating population characteristics, but for predictive inference.

5.4 Prediction regions

In this section, we consider Banks' comparison method for prediction intervals to explore the performance of different bootstrap methods in predictive inference. Here, we intend to investigate the global measure of coverage accuracy for prediction intervals, which are called prediction regions. We then focus on studying how the performance of the PP-B method is affected when the assumed model for PP-B is incorrectly specified. We create the prediction regions using percentile points as we have done in confidence regions. Then, we use a chi-squared test of goodness of fit to assess the discrepancy in coverage probability for the bootstrap prediction intervals. The 20 prediction regions with a nominal coverage probability of 0.05 can be obtained by

$$\begin{aligned} PRL_{(i)} &= \left(q_{(\frac{\alpha_{i+1}}{2})}, q_{(\frac{\alpha_i}{2})}\right) \quad (5.4) \\ PRR_{(i)} &= \left(q_{(1-\frac{\alpha_i}{2})}, q_{(1-\frac{\alpha_{i+1}}{2})}\right) \quad (5.5) \end{aligned}$$

where $i = 1, 2, \ldots, 10$, $\alpha_{i+1} = \alpha_i - 0.10$, $\alpha_1 = 1$ and $q_{(z)}$ is the z^{th} quantile of statistical values, so $PRL_{(i)}$ and $PRR_{(i)}$ are the prediction regions representing the left tail and right tail of the global measure of coverage accuracy, respectively. A total of 10 prediction regions are created, each with a nominal coverage probability of 0.10 can be obtained using Equations (5.4) and (5.5) as follows:

$$PR_{(i)} = PRL_{(i)} \cup PRR_{(i)} \quad (5.6)$$

The following processes are used to study the coverage proportions in the 10 and 20 prediction regions for the future sample statistic based on the bootstrap method:

$PR_{(i)}$	PP-B	NPI-B	PB	EB
1	0.096	0.108	0.073	0.068
2	0.106	0.097	0.071	0.070
3	0.090	0.119	0.072	0.077
4	0.094	0.091	0.076	0.077
5	0.104	0.087	0.078	0.072
6	0.102	0.112	0.072	0.065
7	0.110	0.113	0.090	0.098
8	0.103	0.092	0.100	0.106
9	0.090	0.103	0.144	0.130
10	0.105	0.078	0.224	0.237

Table 5.9: *The coverage proportions for the mean in the 10 prediction regions, where* $n = 50$.

1. Draw a sample $X = (x_1, \ldots, x_n)$ of n observations from a specific distribution to be the past sample and then draw a sample $Y = (y_1, \ldots, y_m)$ of m observations from the same distribution to be the future sample. The samples X and Y are assumed to be independent samples.

2. Compute the statistic of the Y sample, T_m.

3. Draw B bootstrap samples of size m from the X sample and compute the statistic T_m^* for each bootstrap sample to obtain a list of $T_m^*(j)$ for $j = 1, \ldots, B$.

4. Create the 10 and 20 prediction regions for T_m by Equations (5.4), (5.5) and (5.6).

5. Determine if these prediction regions include the statistic T_m.

6. Steps 1-5 are performed in total N times in order to find the coverage proportions.

A simulation study is carried out as described above using different bootstrap methods to compute the coverage proportions in the 10 and 20 prediction regions. Simulations are performed $N = 1000$ times for the mean with a sample size of 50 from Beta(3,1) and the bootstrap methods are applied to each past sample $B = 1000$. The coverage proportions for the mean in the 10 and 20 prediction regions are outlined in Tables 5.9 and 5.10, respectively. The PP-B and NPI-B methods illustrate their superiority in achieving coverage proportions in each of the 10 and 20 prediction regions close to 0.10 and 0.05, respectively. In contrast, the PB and EB methods lead to coverage proportions far from the nominal level of 0.10 in most of the 10 prediction regions, and far from 0.05 in most of the 20 prediction regions. We use the chi-square test to assess the discrepancy between the nominal coverage probabilities and coverage proportions in order to show the best bootstrap method. The resulting χ^2 values are presented in the first row of Table 5.11. The PP-B and NPI-B methods achieve good coverage accuracy, which is reflected by the low chi-squared value in both divisions of the prediction regions. They make the discrepancies between coverage proportions and nominal coverage probabilities lower than the other bootstrap methods. Simulations were repeated several times and consistent results were obtained, as illustrated in Table 5.11. It is obvious that the PP-B method shows its superiority to the other bootstrap methods in achieving the smallest chi-squared values. It distributes the coverage proportions more accurately in most of the prediction region divisions than the other bootstrap methods and this is apparent from having the lowest chi-squared values.

A variety of sample sizes are considered to determine whether the size of the sample affects the performance of different bootstrap methods. Table 5.12 presents the chi-squared values obtained from the coverage proportions for the mean using different bootstrap methods at different sample sizes n. The chi-squared values for all bootstrap methods show no clear pattern as the sample size increases. In both prediction regions, chi-squared values are consistent across all bootstrap methods regardless of sample size. The PP-B method performs better in both prediction region divisions at different sample sizes than any other bootstrap method, followed by NPI-B. For both the PB and the EB methods, the chi-squared values are large because of the great discrepancies between the nominal coverage probabilities and

method	PP-B		NPI-B		PB		EB	
i	$PRL_{(i)}$	$PRR_{(i)}$	$PRL_{(i)}$	$PRR_{(i)}$	$PRL_{(i)}$	$PRR_{(i)}$	$PRL_{(i)}$	$PRR_{(i)}$
1	0.051	0.045	0.057	0.051	0.042	0.031	0.039	0.029
2	0.056	0.050	0.047	0.050	0.040	0.031	0.038	0.032
3	0.043	0.047	0.058	0.061	0.034	0.038	0.038	0.039
4	0.049	0.045	0.042	0.049	0.039	0.037	0.038	0.039
5	0.055	0.049	0.045	0.042	0.042	0.036	0.041	0.031
6	0.051	0.051	0.053	0.059	0.034	0.038	0.030	0.035
7	0.054	0.056	0.055	0.058	0.046	0.044	0.050	0.048
8	0.057	0.046	0.033	0.059	0.050	0.05	0.051	0.055
9	0.044	0.046	0.044	0.059	0.066	0.078	0.060	0.070
10	0.052	0.053	0.028	0.050	0.113	0.111	0.119	0.118

Table 5.10: *The coverage proportions for the mean in the 20 prediction regions, where* $n = 50$.

Repetition	10 PR				20 PR			
	PP-B	NPI-B	PB	EB	PP-B	NPI-B	PB	EB
1	4.42	15.54	216.10	247.00	7.12	30.96	220.36	250.84
2	3.90	9.14	225.62	252.12	13.44	25.16	233.08	259.00
3	5.46	9.34	236.24	253.40	10.88	27.96	241.12	261.44
4	5.48	10.72	234.06	246.24	11.52	27.64	236.24	251.92
5	4.70	11.70	246.48	262.76	11.28	32.64	258.04	265.60
6	4.72	10.42	233.60	241.06	9.80	32.16	238.00	246.12
7	5.34	15.40	227.66	254.52	12.36	32.64	230.64	261.84
8	6.96	10.24	227.04	248.82	18.24	25.56	233.08	253.96
9	6.88	15.76	239.58	238.62	12.48	36.40	243.04	240.52
10	3.18	10.76	227.48	261.34	9.04	33.52	229.96	265.16

Table 5.11: *The chi-squared values obtained from coverage proportions for the mean, where $n = 50$.*

n	measures	10 PR				20 PR			
		PP-B	NPI-B	PB	EB	PP-B	NPI-B	PB	EB
30	χ^2	7.36	16.30	235.78	266.18	25.32	45.60	261.28	287.96
	p-value	0.600	0.061	0.000	0.000	0.150	0.001	0.000	0.000
100	χ^2	7.50	7.56	303.52	314.24	15.32	18.76	306.64	321.48
	p-value	0.585	0.579	0.000	0.000	0.702	0.472	0.000	0.000
200	χ^2	9.66	22.28	281.50	289.32	16.56	31.80	287.28	296.24
	p-value	0.379	0.008	0.000	0.000	0.620	0.033	0.000	0.000
300	χ^2	5.44	6.32	244.30	250.02	13.60	14.48	250.20	252.92
	p-value	0.794	0.708	0.000	0.000	0.806	0.755	0.000	0.000

Table 5.12: *The chi-squared values for the mean and their p-values with different sample sizes n.*

coverage proportions.

We also evaluate different bootstrap methods based on the global accuracy of prediction intervals for the variance. The chi-squared goodness of fit test is used as the basis for comparing the performance of the bootstrap method. We consider several sample sizes to investigate whether or not sample size affects the performance of different bootstrap techniques. In Table 5.13, we present the chi-squared values obtained from the coverage proportions for the variance based on different bootstrap methods. The results of this table are computed in the same manner as before, to demonstrate the performance of these bootstrap techniques. The results of χ^2 values in both prediction region divisions indicate that PP-B and NPI-B are both performing better than any other bootstrap methods. The reason for this is that both methods are able to distribute coverage proportions more accurately across 10 and 20 prediction regions. In contrast, the χ^2 values of PB and EB are high due to the great discrepancies between the nominal coverage probabilities and coverage proportions. The PP-B method has the lowest chi-squared value among these bootstrap methods, which indicates its superiority in achieving coverage proportions close to nominal levels in most of the prediction region divisions.

We investigate how the misspecified PP-B model affects the performance of prediction regions. We consider three different scenarios to evaluate the performance of PP-B in relation to the assumed parametric model. The chi-squared goodness of fit values are computed for the mean and variance with several sample sizes $n = 50, 100, 200, 400$ from an Exponential distribution with rate parameter $\lambda = 5$. In the first scenario, we generate PP-B samples from an Exponential distribution, so the distribution assumed for the PP-B method is well-defined. In the second scenario, the PP-B samples are generated based on the Gamma distribution with parameters α and β. In the third scenario, the PP-B samples are generated using the Normal distribution with parameters μ and σ^2. We generate PP-B samples in the second and third scenarios based on a different distribution from the actual distribution assumed in these data. Table 5.14 shows the chi-squared values obtained from the coverage proportions for the mean and variance under the three scenarios.

The results of χ^2 values for the mean indicate that all different scenarios are able

n	measures	10 PR				20 PR			
		PP-B	NPI-B	PB	EB	PP-B	NPI-B	PB	EB
50	χ^2	16.20	28.82	220.08	327.56	53.20	88.24	232.84	367.44
	p-value	0.063	0.001	0.000	0.000	0.000	0.000	0.000	0.000
100	χ^2	11.56	19.44	257.24	325.08	45.80	53.20	288.64	383.48
	p-value	0.239	0.022	0.000	0.000	0.001	0.000	0.000	0.000
200	χ^2	13.64	15.14	177.42	193.54	23.84	46.04	182.96	205.48
	p-value	0.136	0.087	0.000	0.000	0.005	0.000	0.000	0.000
300	χ^2	7.12	13.02	242.42	268.70	22.32	31.80	267.08	292.52
	p-value	0.625	0.162	0.000	0.000	0.269	0.033	0.000	0.000

Table 5.13: *The chi-squared values for the variance and their p-values with different sample sizes n.*

to distribute the coverage proportions well over the 10 and 20 prediction regions. These results occur for the mean due to the central limit theorem, as discussed in Section 5.2. The central limit theorem (CLT) states that the sampling distribution of the sample mean becomes closer to a normal distribution when the sample size increases irrespective of the distribution from which we sample. However, the difference between the χ^2 values for the variance appears very clearly in the third scenario at both prediction region divisions. The PP-B samples are generated in the second scenario using a distribution that differs from the actual distribution assumed in these data, but they are closely related to one another. Consequently, we achieve lower χ^2 values compared to the third scenario. However, the performance in the first scenario is mostly better when the distribution assumed for the PP-B method is well-defined.

(a) mean

n	measures	10 PR			20 PR		
		PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$	PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$
50	χ^2	23.62	22.98	14.32	37.40	31.80	59.92
	p-value	0.005	0.006	0.111	0.007	0.033	0.000
100	χ^2	14.16	8.16	8.14	21.32	14.44	28.60
	p-value	0.117	0.518	0.520	0.319	0.757	0.073
200	χ^2	6.70	4.46	10.16	15.48	21.68	26.08
	p-value	0.668	0.879	0.338	0.692	0.300	0.128
300	χ^2	8.92	14.26	7.50	21.28	20.04	28.28
	p-value	0.445	0.113	0.585	0.322	0.392	0.078

(b) variance

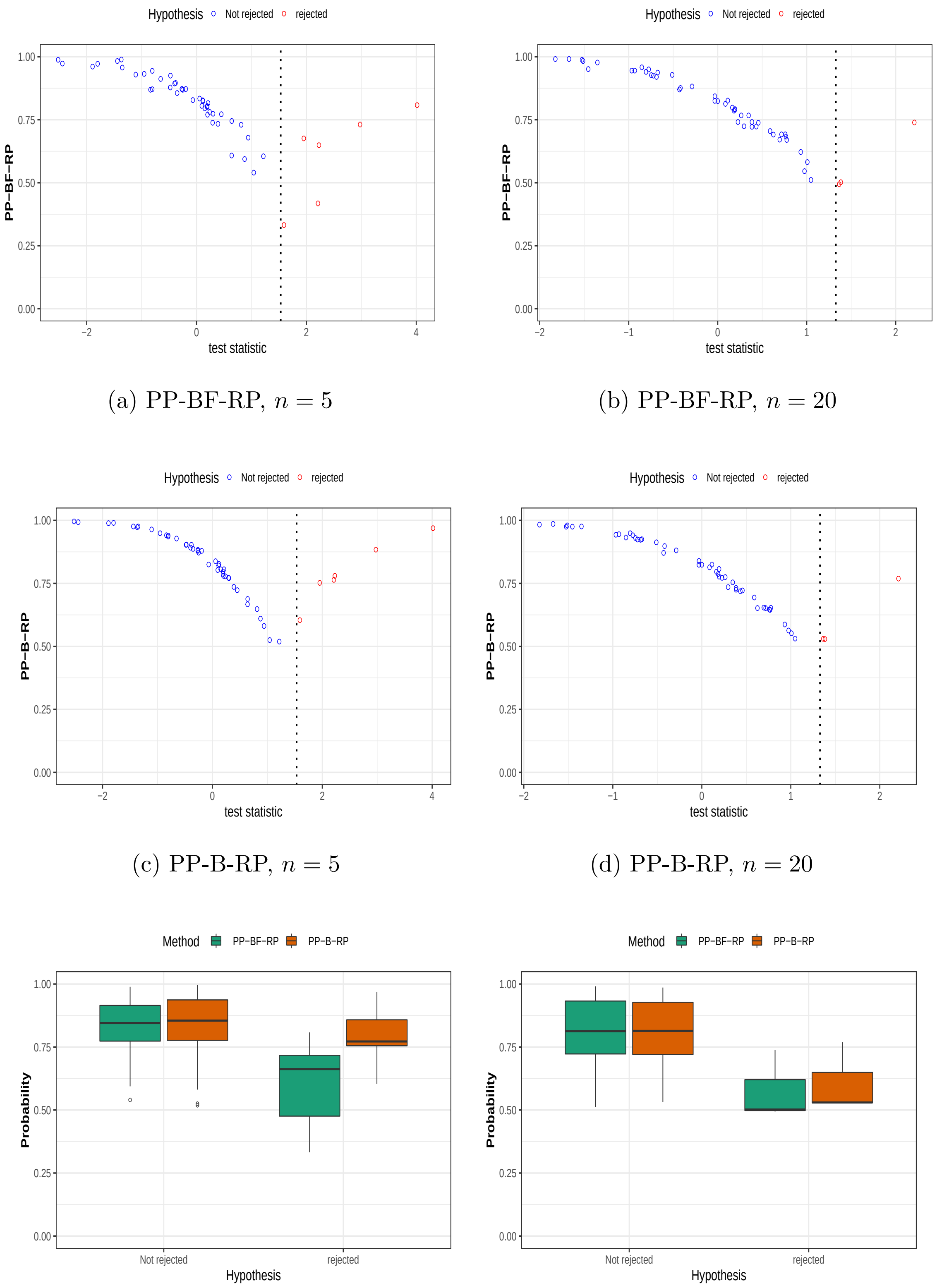
n	measures	10 PR			20 PR		
		PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$	PP-B$_{(1)}$	PP-B$_{(2)}$	PP-B$_{(3)}$
50	χ^2	20.26	64.80	1152.76	36.52	166.56	1202.2
	p-value	0.016	0.000	0.000	0.009	0.000	0.000
100	χ^2	34.90	20.74	1167.04	55.60	76.48	1183.88
	p-value	0.000	0.014	0.000	0.000	0.000	0.000
200	χ^2	6.54	11.34	1016.22	19.44	59.52	1027.8
	p-value	0.685	0.253	0.000	0.429	0.000	0.000
300	χ^2	10.06	9.66	898.74	26.12	55.84	916.20
	p-value	0.346	0.379	0.000	0.127	0.000	0.000

Table 5.14: *The chi-squared values for the mean and variance with their p-values under the three scenarios.*

5.5 PP-B-RP with fixed variance for the one-sample t-test

In this section, we study the reproducibility probability of one-sample t-test using the PP-B method with a fixed variance parameter. The Normal distribution of data is a prerequisite for applying the one-sample t-test. The PP-B method assumes that the data come from a known distribution with unknown parameters. We obtain the PP-B sample by estimating the parameters of the assumed distribution, then drawing one value from the assumed distribution with the estimated parameters from the available data and adding this value to the data. This is continued m times, where each drawn observation is added to the data and the parameters are re-estimated before sampling the next observation. We illustrated in Section 4.3 how to derive the reproducibility probability for the one-sample t-test using the PP-B method. We will explore how the fixed bootstrap variance of PP-B impacts on the RP patterns of one-sample t-test through simulations. The PP-B samples are generated by estimating the mean parameter only from the available data and fixing the variance parameter. The PP-BF-RP is an acronym for the reproducibility value based on fixed bootstrap variance for PP-B. Note that we do not consider the three scenarios when studying the RP of a test, as we may reject the null hypothesis when it is true because we unknowingly reject another wrong assumption. Also, parametric tests require statistical assumptions to apply, e.g. the t-test can only be applied to data that follow a Normal distribution.

The one sided one-sample t-test is considered, $H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$, with level of significance $\alpha = 0.10$. We simulate $N = 50$ samples of size $n = 5$ from the Normal distribution with mean 0 and variance 1 under H_0. The values of RP is determined for each of $N = 50$ simulated sample as explained in Section 4.3. The PP-B samples are generated by fixing the variance parameter to a specific value and estimating the mean parameter from the available data. In these simulations, we compare the RP values when the PP-B variance is fixed at 1 with those without a fixed variance. The same simulated samples are used to calculate the RP value based on PP-BF-RP and PP-B-RP. We also study the impact of increasing sample size to $n = 20$ on PP-BF-RP values for the one-sample t-test. It is important to note that the fixed bootstrap variance for PP-B method and the variance of the simulated samples are the same. Figure 5.3 shows the results of RP values using the PP-B method with fixed and non-fixed bootstrap variance under H_0 for samples of size $n = 5, 20$. The boxplots represent RP values for both methods in cases of rejection



(a) PP-BF-RP, $n = 5$

(b) PP-BF-RP, $n = 20$

(c) PP-B-RP, $n = 5$

(d) PP-B-RP, $n = 20$

(e) RP, $n = 5$

(f) RP, $n = 20$

Figure 5.3: Simulation under H_0: values of the RP of one-sample t-test using PP-B with fixed and non-fixed bootstrap variance, where $n = 5, 20$.

Sample	s^2	Test statistic	n	Test threshold	H_0	PP-BF-RP	PP-B-RP
1	0.216	2.211	5	1.533	R	0.418	0.764
2	0.926	1.953			R	0.676	0.752
3	0.236	1.592			R	0.332	0.604
4	0.683	1.217			NR	0.605	0.519
5	1.114	1.042			NR	0.540	0.525
6	0.517	0.940			NR	0.679	0.580
1	0.949	2.209	20	1.328	R	0.739	0.769
2	0.947	1.383			R	0.502	0.529
3	0.877	1.364			R	0.494	0.530
4	1.279	1.048			NR	0.511	0.53
5	0.873	1.007			NR	0.582	0.552
6	1.186	0.975			NR	0.546	0.563

Table 5.15: *Simulation under H_0: values of RP of one-sample t-test using PP-B with fixed and non-fixed bootstrap variance for six observed samples of sizes $n = 5$ and $n = 20$.*

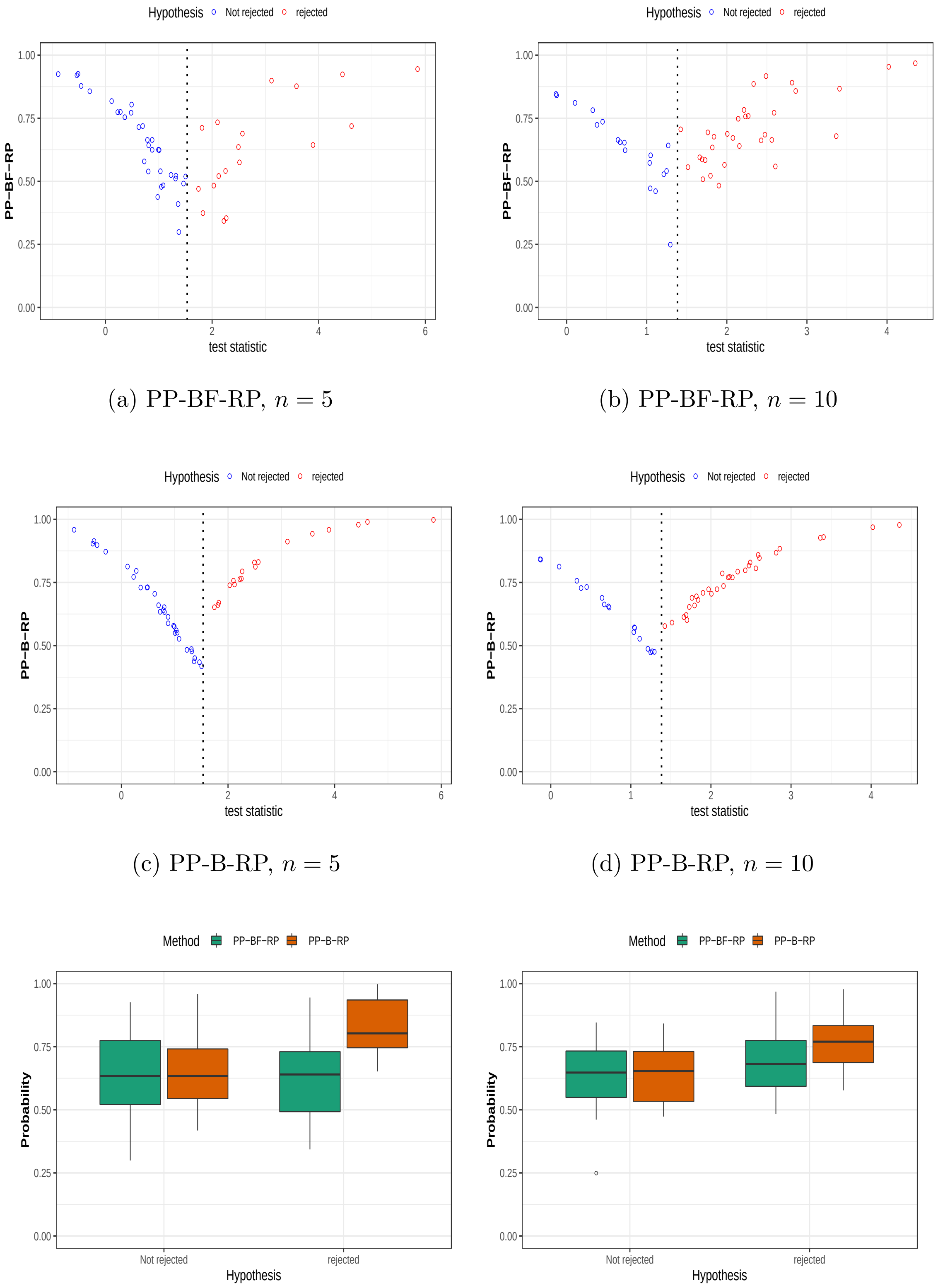
and non-rejection of the null hypothesis. It is interesting to note that there is a high fluctuation of the RP values with fixed bootstrap variance for a small sample of size $n = 5$, particularly when the observed test statistics are close to the test threshold. We observed that increasing the sample size impacts on the patterns of RP values for fixed bootstrap variance. When the sample size increases to $n = 20$, the RP values based on fixed bootstrap variance seem to be similar to those with non-fixed bootstrap variance.

Table 5.15 presents six samples close to the test threshold that reject and do not reject H_0 with samples of sizes $n = 5$ and $n = 20$ for simulations under H_0. For each sample in this table, we present the observed sample variance, test statistics, test threshold, PP-BF-RP and PP-B-RP. The small samples are somewhat more likely to underestimate the population variance as appears from sample variances when

$n = 5$. A larger sample size improves the estimation of the population variance and becomes closer to the true variance of the population as shown when $n = 20$. The PP-BF-RP tends to be lower in the case of rejection when the simulated sample has a small variance such as sample 1 when $n = 5$. In contrast, it tends to be higher in the case of non-rejection when the simulated sample variance is small such as sample 6 when $n = 5$. However, the difference between PP-BF-RP and PP-B-RP methods is reducing as the sample size increases.

The RP is computed by generating B bootstrap samples from the original sample, then applying the one-sample t-test for each of these bootstrap samples. The ratio of B times that have the same decision as the original sample is the RP value. We compute the test statistic of the one-sample t-test using Formula (4.1), which includes the sample variance in the denominator. In the case of rejection, the PP-BF-RP tends to be lower when the variance of the simulated sample is small due to the computed test statistic from these bootstrap samples tending to lie in the non-rejection region. This occurs because the bootstrap sample with fixed variance leads to a smaller test statistic than the test statistic of the simulated sample, as a result of a higher variance value in the denominator. Hence, we obtain more cases that do not reject H_0 due to a test statistic value being smaller than the test threshold. In contrast, PP-BF-RP tends to be higher in the case of non-rejection when the simulated sample variance is small. It is the same reason in the case of rejection, where we obtain more cases of the same decision of an original sample that does not reject H_0. The fluctuation in RP values with fixed bootstrap variance occurs due to a contradiction between the simulated sample variance and bootstrap sample variance.

The simulations are performed assuming the underlying populations under the alternative hypothesis, $H_a : \mu > \mu_0$, with level of significance $\alpha = 0.10$. We simulate $N = 50$ samples of sizes $n = 5, 10$ from the Normal distribution with mean 0.5 and variance 1 under H_a. The values of RP are determined for each of $N = 50$ simulated samples using PP-B with fixed and non-fixed variance. We compare the RP values when the PP-B variance is fixed at 1 with the non-fixed bootstrap variance. The same data sets for each sample are used to compute the RP value based on PP-B



(a) PP-BF-RP, $n = 5$

(b) PP-BF-RP, $n = 10$

(c) PP-B-RP, $n = 5$

(d) PP-B-RP, $n = 10$

(e) RP, $n = 5$

(f) RP, $n = 10$

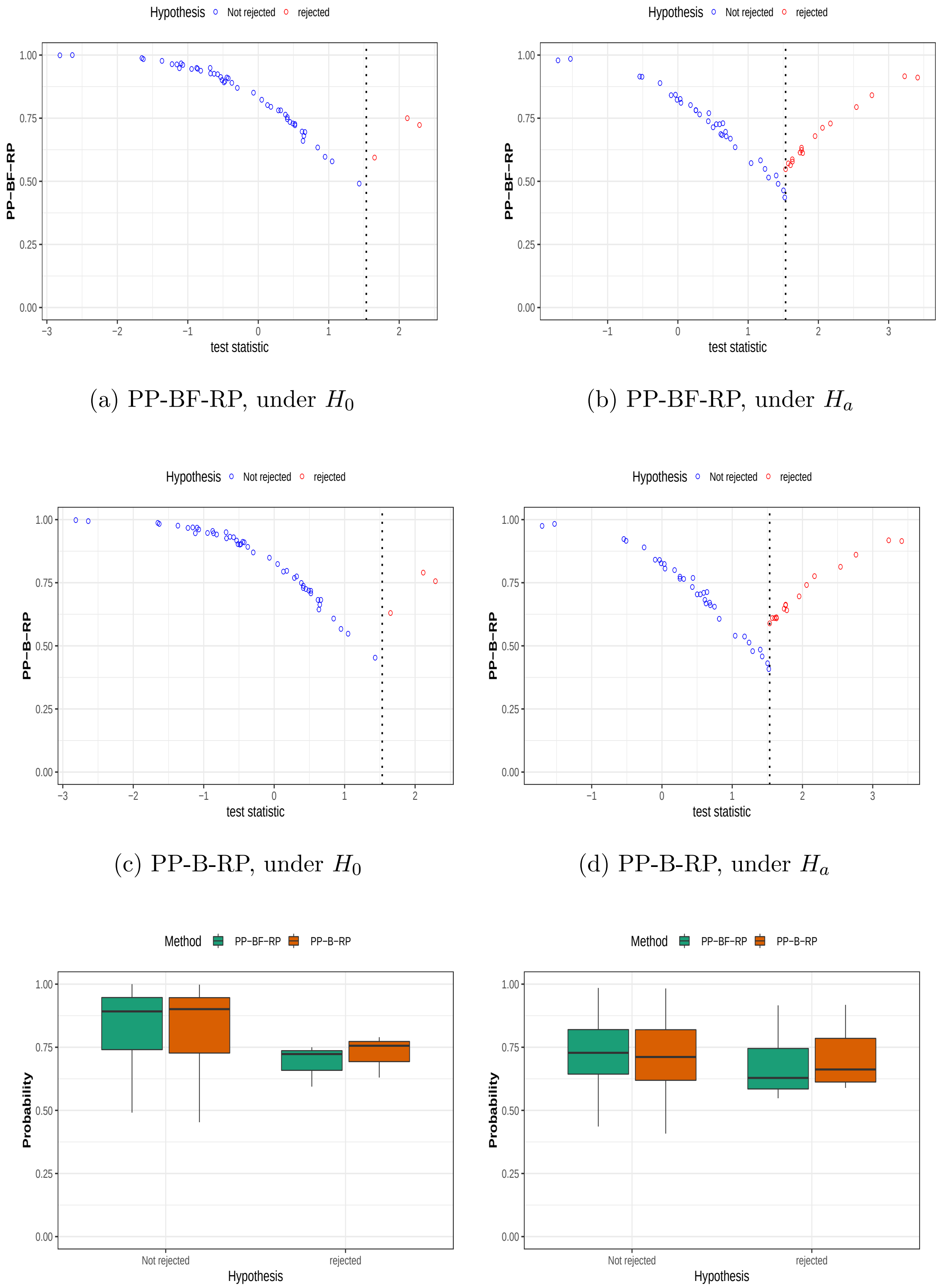
Figure 5.4: Simulation under H_a: values of the RP of one-sample t-test using PP-B with fixed and non-fixed bootstrap variance, where $n = 5, 10$.

Sample	s^2	Test statistic	n	Test threshold	H_0	PP-BF-RP	PP-B-RP
1	0.263	1.828	5	1.533	R	0.374	0.670
2	1.427	1.811			R	0.712	0.661
3	0.531	1.744			R	0.470	0.652
4	0.727	1.504			NR	0.519	0.418
5	0.924	1.464			NR	0.491	0.434
6	2.310	1.376			NR	0.299	0.450
1	1.662	1.662	10	1.328	R	0.595	0.613
2	0.962	1.514			R	0.556	0.591
3	1.941	1.424			R	0.706	0.577
4	2.759	1.294			NR	0.249	0.475
5	0.396	1.267			NR	0.642	0.477
6	0.789	1.247			NR	0.541	0.473

Table 5.16: *Simulation under H_a: values of RP of one-sample t-test using PP-B with fixed and non-fixed bootstrap variance for six observed samples of sizes $n = 5$ and $n = 10$.*

with fixed and non-fixed variance. The results of RP values based on the PP-BF-RP and PP-B-RP methods with samples of size $n = 5, 10$ under H_a are presented in Figure 5.4. The boxplots of RP values are displayed for the two methods in both cases of rejection and non-rejection. The fluctuation in RP values with fixed bootstrap variance is more visible when simulations are conducted under H_a due to more cases of test statistics close to the test threshold.

A total of six samples close to the test threshold that reject and do not reject H_0 with samples of sizes $n = 5$ and $n = 20$ for simulations under H_a are shown in Table 5.16. From this table, we see that some simulated samples overestimate the population variance. When the simulated sample has a large variance, the PP-BF-RP is completely opposite the small variance of the simulated sample. In the case of rejection, the PP-BF-RP tends to be higher when the simulated sample variance



(a) PP-BF-RP, under H_0

(b) PP-BF-RP, under H_a

(c) PP-B-RP, under H_0

(d) PP-B-RP, under H_a

(e) RP, under H_0

(f) RP, under H_a

Figure 5.5: Simulation under H_0 and H_a: values of the RP of one-sample t-test using PP-B with fixed and non-fixed bootstrap variance, where $n = 5$ and variances of all simulated samples lie between 0.98 and 1.02.

is large, such as in sample 3 when $n = 10$. Conversely, it has a lower RP value in the case of non-rejection when the variance of the simulated sample is large such as sample 4 when $n = 10$. In the case of the simulated sample with large variance, the bootstrap sample with a fixed variance of 1 leads to larger test statistics, as a result of a smaller variance value in the denominator. Hence, we obtain more cases that reject H_0 due to a test statistic value being larger than the test threshold.

In the following simulations, we repeat the same procedure to compare the RP values for a one-sample t-test using PP-B with fixed and non-fixed variance. The simulations are conducted under H_0 and H_a for samples of size $n = 5$, but with restricted variances of all simulated samples between 0.98 and 1.02. Consequently, the variances of all simulated samples are very close to the fixed bootstrap variance. Figure 5.5 shows the simulation results of both PP-BF-RP and PP-B-RP methods under H_0 for samples of size $n = 5$. The boxplots of RP values are presented for both methods in cases of rejection and non-rejection of the null hypothesis. It is obvious that both methods produce similar RP values due to the variances of the simulated samples and bootstrap samples are close. This approves the fluctuation in PP-BF-RP values occurring because of a discrepancy between the variances of the simulated sample and bootstrap sample.

5.6 Concluding remarks

This chapter examines the impact of the misspecified PP-B model in a range of scenarios, as well as how the fixed bootstrap variance of PP-B affects RP for one-sample t-test. Misspecification of models can result in several undesirable consequences and may lead to unexpected behaviour. A misspecified PP-B model refers to situations in which the distribution of data sets used is incorrectly specified. The performance of PP-B is examined under three different scenarios regarding the assumed parametric model. The first scenario generates PP-B samples from the same distribution as the original data set, indicating that the distribution assumed for the PP-B method is well-defined. The second scenario generates PP-B samples from a different distribution than that used to sample the original data set, but they are closely related to one another. The third scenario generates PP-B samples from a completely different distribution than that used for the original data set.

We investigate the influence of an incorrectly specified PP-B model on the performance of the LC prediction interval through simulation studies. All different scenarios for the mean provide coverage that is close to the nominal coverage probability due to the central limit theorem. The difference appears clearly in the third scenario for the variance, which performed poorly with under-coverage far from the nominal coverage probabilities. Banks' comparison strategy of the bootstrap confidence interval is used to compare PP-B with other bootstrap methods. PP-B and NPI-B provide wide confidence regions that arise from greater variability in their bootstrap samples, so they have over-coverage in many confidence regions. As a result, the chi-squared value for these bootstrap methods is high due to the large discrepancy between the nominal coverage probabilities and coverage proportions. In contrast, the PB and EB methods perform well in achieving low discrepancies between the nominal coverage probabilities and coverage proportions.

We illustrate how to apply Banks' comparison method for the bootstrap prediction interval, which is called prediction regions. The performance of PP-B is compared to other bootstrap methods, as well as the impact of the misspecified PP-B model on the prediction region's performance. The PP-B method performs best with the lowest chi-squared values, followed by the NPI-B method. These bootstrap methods are able to distribute coverage proportions more accurately across prediction regions divisions. The chi-squared values of PB and EB methods are large because of the great discrepancies between the nominal coverage probabilities and coverage proportions. The performance of prediction regions is examined in relation to the misspecified PP-B model. The results of chi-squared values for the mean show that all different scenarios are able to distribute the coverage proportions well across prediction regions divisions due to the central limit theorem. The difference between the χ^2 values appears clearly for the variance, in particular for the third scenario. The PP-B samples are generated in the second scenario using a different distribution than that used for the original data set, but they are closely related to one another. Consequently, the PP-B method performs much better in the second scenario compared to the third scenario. However, the performance of the PP-B method is mostly better when the distribution assumed for the PP-B method is well-defined. Finally, we investigate the reproducibility probability of one-sample t-test using PP-B with a fixed variance parameter. The fixed bootstrap variance of PP-B has a negative impact on the RP value for one-sample t-test. This occurs because of violating the RP definition which states an experiment is performed in the same way as the original experiment.

Chapter 6

Conclusions

In this chapter, we summarize the main results of this thesis and conclude with some future research topics. A new version of bootstrap is presented in this thesis, namely parametric predictive bootstrap (PP-B). It has been applied in a range of scenarios in order to evaluate its performance with other bootstrap methods. The PP-B method is used to predict the reproducibility probability of hypothesis tests.

In Chapter 3, we presented the main concept of the parametric predictive bootstrap. The procedure depends on estimating the parameters of the assumed distribution from the original data set of size n, then drawing one observation from the assumed distribution based on the estimated parameters and adding it to the original data. So, the first sampled observation is added to the data set, leading to $n + 1$ observations. The second observation is then drawn from the same underlying distribution with estimated parameters anew from $n + 1$ observations. This is continued to sample m further values in the same manner, each one adding to the data and re-estimating parameters before sampling the next one. These sampled values represent the observations of PP-B which of course does not include the n original data observations. As a result of the method of sampling observations, PP-B has more variation than PB and EB. The PP-B method does not perform well in estimation using some measures of statistical accuracy and confidence intervals. PP-B has generally greater values of statistical accuracy measures for estimators. Also, it has an over-coverage tendency for percentile confidence intervals due to wide intervals arising from greater variability in their bootstrap samples. However, it performs very well with prediction when we test its performance with the LC percentile prediction interval. An advantage of the PP-B method is that it gives a good coverage probability with a narrower average width of intervals compared to the NPI-B method, as shown in our simulation studies.

In Chapter 4, we discussed the PP-B method of the RP of some parametric tests (PP-B-RP). It has been noted that, there is no single definition of RP within the classical frequentist statistics framework. We consider the main idea of RP, which is how likely it is that the same test result would be obtained if the experiment were repeated under identical conditions as the original experiment. Test reproducibility is more naturally viewed as a prediction problem than as an estimation problem, which is in line with the PP-B's predictive nature. Simulation studies are used to compare the RP of tests based on PP-B with a similar predictive bootstrap method, NPI-B. As a result of the assumption of a parametric model in PP-B, RP values based on PP-B have less variability than those obtained with NPI-B. We found the PP-B-RP method tends to provide a value within the lower and upper NPI-RP for the likelihood ratio test, meaning this bootstrap method is consistent with the NPI-RP approach. The employment of the bootstrap approach with RP has the advantage of avoiding the complexities involved in computations of the lower and upper bounds in NPI-RP.

In Chapter 5, we study the effect of the misspecified PP-B model in a range of scenarios. Also, we explore how the fixed bootstrap variance of PP-B impacts on the value of RP for one-sample t-test. Three different scenarios are investigated to evaluate the performance of PP-B regarding the assumed parametric model. In the first scenario, the model used in PP-B is the same model assumed in the original data set. The second and third scenarios illustrate how PP-B performs when the model assumed for PP-B is different from the actual model assumed in the original data. The distribution assumed for PP-B is closely related to the distribution of the original data in the second scenario, but it is a completely different distribution than that used for the original data set in the third scenario. All different scenarios for the mean provide similar performance as a result of the central limit theorem. The difference appears clearly in the third scenario for variance because the model assumed for PP-B is completely different from the actual model assumed to sample the original data. The PP-B method does not perform well in confidence regions of the bootstrap confidence interval, as it is not developed for estimating population characteristics. It is explicitly aimed at predictive inference, with variability in different bootstrap samples reflecting uncertainty. The PP-B method performs best in prediction regions, as evidenced by achieving the lowest chi-squared values. The large chi-squared values refer to great discrepancies between the nominal coverage probabilities and coverage proportions. The PP-B method is able to distribute the coverage proportions more accurately in most of the prediction region divisions than the other bootstrap methods. It is apparent that the fixed bootstrap variance of PP-B adversely affected the RP value of one-sample t-test. The reason is that it violates the definition of RP by performing an experiment differently from the original experiment.

Many future research directions can be explored based on the work presented in this thesis. The PP-B method was developed by starting with only one dimension of real-valued data, but it can be extended to include two or more dimensions. One important topic for future work is to study the influence of outliers in the data on the proposed PP-B method. We present the PP-B method to determine the reproducibility probability of some parametric tests. This method can be applied to a wide range of parametric statistical tests and extended in different ways, such as using future sample sizes that are different from the data sample size or using different significance levels. Bootstrap methods have been widely used in many statistical situations for precise inferences. They may be extended for imprecise inferences to provide lower and upper bounds rather than just one single value. Future research into the PP-B method can be based on these suggestions.

Appendix A

Confidence intervals

We use the BCa interval in Chapter 3 to evaluate the performance of different bootstrap methods as an estimation approach. Simulation studies are conducted to find the coverage proportion and average width of intervals for three statistics: mean, variance, and median. We use the Exponential distribution with rate parameter λ to generate data. The probability density function of the Exponential distribution is as follows

$$f(x) = \lambda e^{-\lambda x} \quad ; \quad x \in [0, \infty) \tag{A.1}$$

The original sample size n is generated from Exp(4), then different bootstrap methods are applied $B = 1000$ times. For each bootstrap sample, we compute the statistics in order to construct BCa intervals based on Equation (2.27). Following this, we identify which BCa confidence intervals include the true statistics of the Exp(4) distribution. We repeated this procedure $N = 1000$ times to find the coverage proportions of different bootstrap methods with different original sample sizes $n = 50, 100, 200, 400$ and confidence levels 95% and 90%. Table A.1 shows the coverage proportions and average interval widths for several statistics based on the four bootstrap procedures. Note that the bootstrap samples for each method are the same size as the original samples.

PP-B and NPI-B intervals lead to wider confidence intervals than the other bootstrap methods due to the greater variability in their bootstrap samples. Consequently, the NPI-B method has over-coverage in all cases of the three statistics, as well as for the mean in the PP-B method. In the case of the mean, the PB and EB

(a) mean

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.9910	0.9970	0.9960	0.9950	0.9740	0.9850	0.9770	0.9770
	AW	0.2100	0.1437	0.0994	0.0698	0.1737	0.1196	0.0832	0.0585
NPI-B	CP	0.9920	0.9960	0.9940	0.9960	0.9730	0.9850	0.9760	0.9770
	AW	0.2178	0.1482	0.1019	0.0709	0.1770	0.1222	0.0849	0.0592
PB	CP	0.9470	0.9570	0.9440	0.9490	0.9020	0.9080	0.9040	0.8980
	AW	0.1429	0.0994	0.0696	0.0491	0.1193	0.0833	0.0584	0.0412
EB	CP	0.9330	0.9500	0.9360	0.9460	0.8850	0.9100	0.9050	0.8940
	AW	0.1386	0.0980	0.0691	0.0488	0.1160	0.0822	0.0580	0.0410

(b) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.8730	0.8780	0.8780	0.8790	0.8030	0.8130	0.8130	0.8100
	AW	0.1563	0.0985	0.0645	0.0430	0.1313	0.0826	0.0546	0.0365
NPI-B	CP	0.9870	0.9880	0.9910	0.9930	0.9460	0.9720	0.9740	0.9750
	AW	0.2503	0.1506	0.0950	0.0615	0.1569	0.1037	0.0696	0.0472
PB	CP	0.7940	0.7980	0.8080	0.7910	0.7220	0.7260	0.7190	0.7040
	AW	0.1156	0.0744	0.0499	0.0335	0.0974	0.0627	0.0423	0.0285
EB	CP	0.8460	0.9070	0.9240	0.9210	0.7900	0.8500	0.8700	0.8730
	AW	0.0909	0.0680	0.0493	0.0353	0.0764	0.0566	0.0411	0.0293

(c) median

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	CP	0.8680	0.8590	0.8460	0.8570	0.7820	0.7920	0.7650	0.7830
	AW	0.1667	0.1153	0.0804	0.0568	0.1416	0.0983	0.0684	0.0484
NPI-B	CP	0.9960	0.9940	0.9940	0.9940	0.9820	0.9790	0.9820	0.9750
	AW	0.1997	0.1402	0.0983	0.0690	0.1667	0.1173	0.0825	0.0581
PB	CP	0.7700	0.7750	0.7510	0.7700	0.6860	0.6790	0.6850	0.6870
	AW	0.1274	0.0896	0.0633	0.0450	0.1092	0.0768	0.0540	0.0385
EB	CP	0.9420	0.9560	0.9500	0.9410	0.8990	0.9070	0.8890	0.8880
	AW	0.1355	0.0972	0.0685	0.0485	0.1135	0.0817	0.0572	0.0407

Table A.1: *Coverage of* $100(1-2\alpha)\%$ *confidence interval using BCa for different statistics when the original sample from Exp(4).*

have good coverage with smaller average interval widths. EB provide the nominal coverage probability 0.95 for the mean when $n = 100$. The coverage proportions of the PP-B for variance and median are better than the PB, but both methods show under-coverage results in all cases. It is interesting to note that PP-B sometimes shows under-coverage results despite its wide intervals with the BCa method. PP-B and PB produce large values of bias-correction when the BCa method is used for the variance and median. These large values have adversely affected the BCa interval endpoints. The EB achieves better coverage proportions than other bootstrap methods for most cases of variance and all cases of median. However, it shows worse undercoverage results for variance when sample sizes are small. For example, the coverage proportions of EB are almost 11% below their 95% and 90% nominal confidence levels when $n = 50$.

We study PP-B and PB with the BC interval to discover if the bias-correction value is indeed responsible for the under-coverage of the variance and median with Exp(4). We present the results of coverage and average interval widths using PP-B and PB with the BC method in Table A.2. The same original sample sizes are used with the BC and BCa methods, also the same seeds are also applied to all bootstrap methods. The median results for PP-B and PB using the BC method are identical to those using the BCa method in Table A.1 because the acceleration value for the median is zero, but they are different for the variance. Coverage of the variance is slightly increased in some cases with the BC method, but they are still far from nominal coverage probabilities.

We examine the performance of different bootstrap methods using the percentile interval in Equation (2.17) and compare it with the BCa interval. We conduct simulations using the same original samples of Exp(4) that were used in the BCa interval studies. The results of coverage and average interval widths for the mean, variance, and median with Exp(4) are shown in Table A.3. There is a considerable difference between the two interval methods with PP-B and PB when comparing variance and median with Exp(4). The BCa method provides under-coverage results in all cases of variance and median. In contrast, the percentile method has over-coverage with PP-B and PB in all cases of variance and median. The percentile

(a) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	*CP*	0.8740	0.8790	0.8790	0.8730	0.8060	0.8140	0.8190	0.8120
	AW	0.1344	0.0884	0.0603	0.0413	0.1125	0.0743	0.0511	0.0351
PB	*CP*	0.8000	0.8040	0.8080	0.7930	0.7090	0.7180	0.7220	0.7110
	AW	0.1021	0.0682	0.0471	0.0324	0.0858	0.0576	0.0402	0.0277

(b) median

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	*CP*	0.8680	0.8590	0.8460	0.8570	0.7820	0.7920	0.7650	0.7830
	AW	0.1667	0.1153	0.0804	0.0568	0.1416	0.0983	0.0684	0.0484
PB	*CP*	0.7700	0.7750	0.7510	0.7700	0.6860	0.6790	0.6850	0.6870
	AW	0.1274	0.0896	0.0633	0.0450	0.1092	0.0768	0.0540	0.0385

Table A.2: *Coverage of* $100(1-2\alpha)\%$ *confidence interval using BC method for the variance and median with original sample from Exp(4).*

method achieved similar results as the BCa method for the mean with all bootstrap methods, also for variance and median when using NPI-B and EB. It occurs due to the BCa method having smaller bias-correction and acceleration values, which leads to the endpoints of the BCa interval being close to the percentile interval.

(a) mean

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	*CP*	0.9900	0.9960	0.9930	0.9950	0.9720	0.9830	0.9760	0.9790
	AW	0.1942	0.1376	0.0973	0.0690	0.1617	0.1153	0.0816	0.0580
NPI-B	*CP*	0.9920	0.9960	0.9970	0.9980	0.9770	0.9830	0.9800	0.9830
	AW	0.2227	0.1513	0.1036	0.0717	0.1831	0.1255	0.0864	0.0600
PB	*CP*	0.9350	0.9570	0.9460	0.9500	0.8910	0.9100	0.8980	0.8980
	AW	0.1384	0.0977	0.0690	0.0489	0.1161	0.0821	0.0580	0.0411
EB	*CP*	0.9260	0.9470	0.9410	0.9460	0.8860	0.9010	0.9030	0.8880
	AW	0.1344	0.0963	0.0685	0.0485	0.1131	0.0810	0.0576	0.0408

(b) variance

Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	*CP*	0.9950	0.9990	0.9990	0.9990	0.9810	0.9970	0.9930	0.9960
	AW	0.1187	0.0841	0.0594	0.0422	0.0956	0.0691	0.0493	0.0353
NPI-B	*CP*	0.9940	0.9920	0.9950	0.9960	0.9720	0.9760	0.9790	0.9810
	AW	0.2194	0.1428	0.0948	0.0630	0.1602	0.1085	0.0739	0.0501
PB	*CP*	0.9850	0.9970	0.9940	0.9940	0.9660	0.9790	0.9720	0.9760
	AW	0.0985	0.0693	0.0487	0.0345	0.0802	0.0572	0.0406	0.0289
EB	*CP*	0.8160	0.8780	0.9060	0.9160	0.7630	0.8150	0.8600	0.8620
	AW	0.0760	0.0590	0.0444	0.0327	0.0654	0.0501	0.0376	0.0275

(c) median

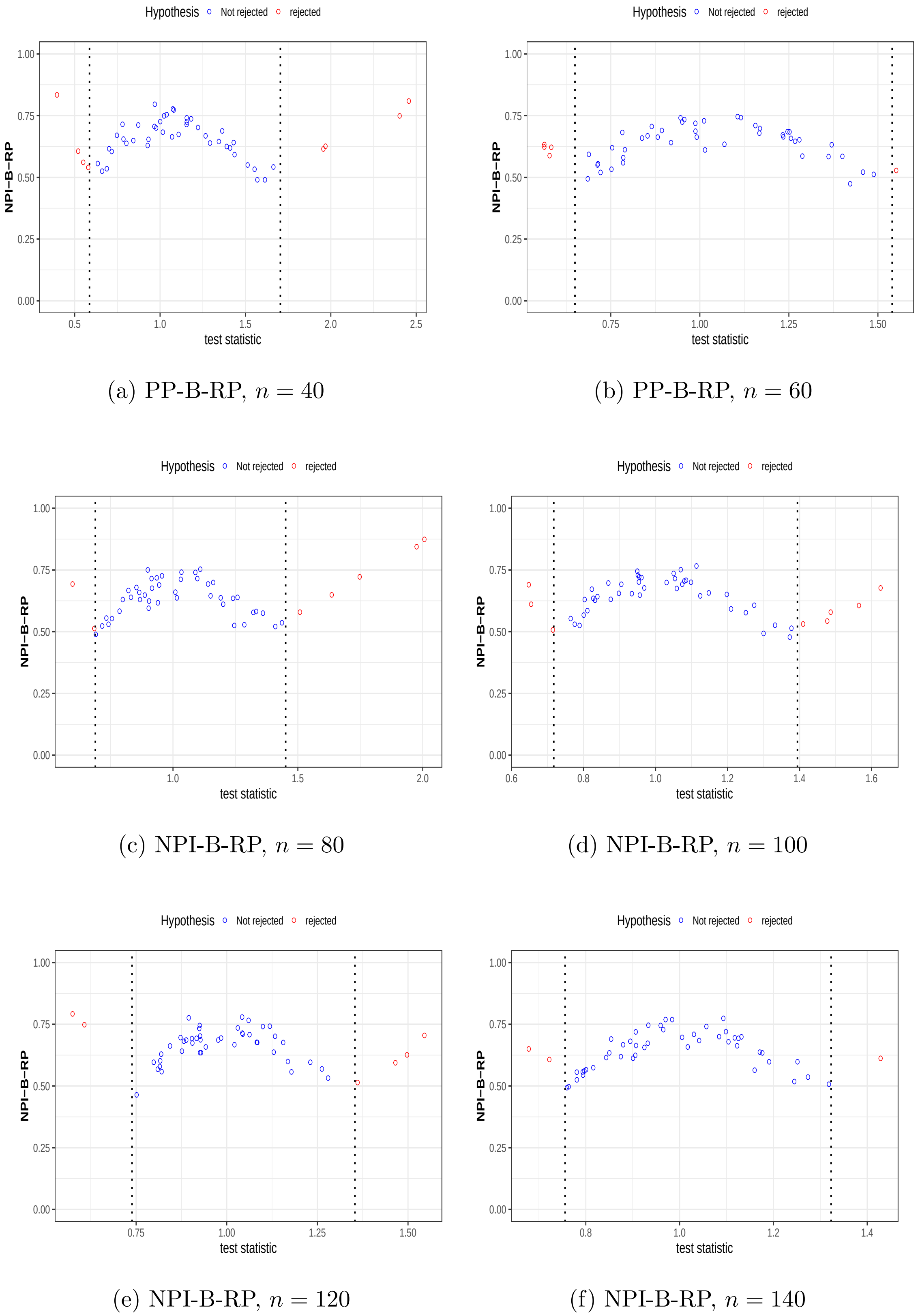
Bootstrap	measures	Confidence level							
		95%				90%			
		$n = 50$	$n = 100$	$n = 200$	$n = 400$	$n = 50$	$n = 100$	$n = 200$	$n = 400$
PP-B	*CP*	0.9990	1.0000	1.0000	0.9990	0.9910	0.9970	0.9960	0.9970
	AW	0.1700	0.1196	0.0840	0.0595	0.1415	0.1000	0.0705	0.0500
NPI-B	*CP*	0.9970	0.9950	0.9940	0.9960	0.9790	0.9830	0.9830	0.9740
	AW	0.1997	0.1403	0.0982	0.0691	0.1667	0.1174	0.0825	0.0581
PB	*CP*	0.9900	0.9970	0.9960	0.9970	0.9740	0.9880	0.9760	0.9810
	AW	0.1375	0.0974	0.0688	0.0488	0.1152	0.0818	0.0578	0.0410
EB	*CP*	0.9430	0.9550	0.9430	0.9450	0.9050	0.9090	0.8940	0.8840
	AW	0.1383	0.0979	0.0688	0.0486	0.1159	0.0823	0.0575	0.0408

Table A.3: *Coverage of* $100(1-2\alpha)\%$ *confidence interval using percentile method for different statistics with original sample from Exp(4).*

Appendix B

NPI-B-RP for the F-test

We study the RP of the F-test using NPI-B with a variety of sample sizes. This helps us to observe how the NPI-B method performs for the RP of the F-test as the size of samples increases. The two sided F-test is considered, $H_0 : \sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$, and level of significance $\alpha = 0.10$. We simulate two samples of size n under H_0 in total $N = 50$ times. The data are generated for the two original samples from the same Normal distribution with mean 0 and standard deviation 1. In Section 4.5, we discussed how to determine the RP of the F-test based on the bootstrap method. We perform simulation studies for two samples of size $n = 40, 60, 80, 120, 140$ to explore whether there is an influence of sample size on the variability of NPI-B-RP values for the F-test. The observed test statistic and the Bootstrap-RP were determined for each of the $N = 50$ samples. It is important to note that the bootstrap samples for each method have the same size as the original sample. Figure B.1 shows the results of RP values using NPI-B methods under H_0 with a variety of sample sizes. Simulation studies show that the NPI-B-RP values fluctuate clearly even with increasing the size of samples.



(a) PP-B-RP, $n = 40$

(b) PP-B-RP, $n = 60$

(c) NPI-B-RP, $n = 80$

(d) NPI-B-RP, $n = 100$

(e) NPI-B-RP, $n = 120$

(f) NPI-B-RP, $n = 140$

Figure B.1: Simulations under H_0: values of NPI-B-RP for F-test.

Bibliography

[1] Ahad N.A. and Yahaya S.S.S. (2014). Sensitivity analysis of Welch'st-test. In *AIP Conference Proceedings*, volume 1605, pp. 888–893. American Institute of Physics.

[2] Al Luhayb A.S.M., Coolen F.P.A. and Coolen-Maturi T. (2019). Generalizing Banks' smoothed bootstrap method for right-censored data. pp. 894–901. 29th European Safety and Reliability Conference (ESREL 2019), Hannover (Germany).

[3] Augustin T. and Coolen F.P.A. (2004). Nonparametric predictive inference and interval probability. *Journal of Statistical Planning and Inference*, 124(2), 251–272.

[4] Banks D.L. (1988). Histospline smoothing the Bayesian bootstrap. *Biometrika*, 75(4), 673–684.

[5] Begley C.G. and Ellis L.M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391), 531–533.

[6] Berrar D. (2019). Introduction to the non-parametric bootstrap. In *Encyclopedia of bioinformatics and computational biology*, pp. 766–773. Academic Press Oxford.

[7] Chan W., Yung Y.F., Bentler P.M. and Tang M.L. (1998). Tests of independence for ordinal data using bootstrap. *Educational and Psychological Measurement*, 58(2), 221–240.

[8] Chernick M.R. (2011). *Bootstrap Methods: A Guide for Practitioners and Researchers.* John Wiley & Sons.

[9] Chernick M.R. and Labudde R.A. (2009). Revisiting qualms about bootstrap confidence intervals. *American Journal of Mathematical and Management Sciences*, 29(3-4), 437–456.

[10] Chernick M.R. and LaBudde R.A. (2014). *An Introduction to Bootstrap Methods with Applications to R.* John Wiley & Sons.

[11] Coolen F.P.A. (2006). On nonparametric predictive inference and objective Bayesianism. *Journal of Logic, Language and Information*, 15(1-2), 21–47.

[12] Coolen F.P.A. (2011). Nonparametric predictive inference. In *International Encyclopedia of Statistical Sciences*, (Editor) M. Lovric, pp. 968–970. Berlin: Springer.

[13] Coolen F.P.A. and Alqifari H.N. (2018). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *REVSTAT: Statistical Journal.*, 16(2), 167–185.

[14] Coolen F.P.A. and Elsaeiti M.A. (2009). Nonparametric predictive methods for acceptance sampling. *Journal of Statistical Theory and Practice*, 3(4), 907–921.

[15] Coolen F.P.A. and Himd S.B. (2014). Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, 8(4), 591–618.

[16] Coolen F.P.A. and Himd S.B. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov–Smirnov test. *Journal of Statistical Theory and Practice*, 14(2), 1–13.

[17] Coolen F.P.A. and Marques F.J. (2020). Nonparametric Predictive Inference for Test Reproducibility by Sampling Future Data Orderings. *Journal of Statistical Theory and Practice*, 14(4), 1–22.

[18] Coolen F.P.A., Troffaes M.C.M. and Augustin T. (2011). *Imprecise Probability*, pp. 645–648. Berlin, Heidelberg: Springer Berlin Heidelberg.

[19] Coolen-Maturi T., Coolen-Schrijner P. and Coolen F.P. (2012). Nonparametric predictive inference for binary diagnostic tests. *Journal of Statistical Theory and Practice*, 6, 665–680.

[20] Coolen-Schrijner P., Maturi T.A. and Coolen F.P.A. (2009). Nonparametric predictive precedence testing for two groups. *Journal of Statistical Theory and Practice*, 3(1), 273–287.

[21] Davison A.C. and Hinkley D.V. (1997). *Bootstrap methods and their application.* Cambridge University Press.

[22] De Capitani L. and De Martini D. (2011). On stochastic orderings of the Wilcoxon rank sum test statistic—with applications to reproducibility probability estimation testing. *Statistics & Probability Letters*, 81(8), 937–946.

[23] De Capitani L. and De Martini D. (2015). Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *Journal of Statistical Computation and Simulation*, 85(3), 468–493.

[24] De Capitani L. and De Martini D. (2016). Reproducibility probability estimation and RP-testing for some nonparametric tests. *Entropy*, 18(4), 142.

[25] De Martini D. (2008). Reproducibility probability estimation for testing statistical hypotheses. *Statistics & Probability Letters*, 78(9), 1056–1061.

[26] Delacre M., Lakens D. and Leys C. (2017). Why psychologists should by default use Welch's t-test instead of Student's t-test. *International Review of Social Psychology*, 30(1).

[27] Derrick B., Toher D. and White P. (2016). Why Welch's test is Type I error robust. *The Quantitative Methods in Psychology*, 12(1).

[28] Diaconis P. and Efron B. (1983). Computer-intensive methods in statistics. *Scientific American*, 248(5), 116–131.

[29] DiCiccio T.J. and Efron B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3), 189–228.

[30] Efron B. (1979). Bootstrap methods: another look at the jackknife. *The Annals of Statistics*, 7, 1–26.

[31] Efron B. (1981). Nonparametric standard errors and confidence intervals. *Canadian Journal of Statistics*, 9(2), 139–158.

[32] Efron B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Society for Industrial and Applied Mathematics.

[33] Efron B. (1987). Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397), 171–185.

[34] Efron B. (1990). More efficient bootstrap computations. *Journal of the American Statistical Association*, 85(409), 79–89.

[35] Efron B. and Gong G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36–48.

[36] Efron B. and Hastie T. (2016). *Computer Age Statistical Inference*. Cambridge University Press.

[37] Efron B. and Narasimhan B. (2020). The automatic construction of bootstrap confidence intervals. *Journal of Computational and Graphical Statistics*, (just-accepted), 1–32.

[38] Efron B. and Tibshirani R. (1985). The bootstrap method for assessing statistical accuracy. *Behaviormetrika*, 12(17), 1–35.

[39] Efron B. and Tibshirani R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, pp. 54–75.

[40] Efron B. and Tibshirani R.J. (1994). *An introduction to the bootstrap*. Boca Raton, Florida : Chapman & Hall/CRCl.

[41] Gerald B. (2018). A brief review of independent, dependent and one sample t-test. *International Journal of Applied Mathematics and Theoretical Physics*, 4(2), 50–54.

[42] Goodman S.N. (1992). A comment on replication, P-values and evidence. *Statistics in Medicine*, 11(7), 875–879.

[43] Hall P. (1988). Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, pp. 927–953.

[44] Hesterberg T. (2011). Bootstrap. *Wiley Interdisciplinary Reviews: Computational Statistics*, 3(6), 497–526.

[45] Hill B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322), 677–691.

[46] Hill B.M. (1988). De Finetti's theorem, induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). *Bayesian Statistics*, 3, 211–241.

[47] Hill B.M. (1993). Parametric models for An: splitting processes and mixtures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2), 423–433.

[48] Hosken D., Buss D. and Hodgson D. (2018). Beware the F test (or, how to compare variances). *Animal Behaviour*, 136, 119–126.

[49] Islam M.R. (2018). Sample size and its role in Central Limit Theorem (CLT). *Computational and Applied Mathematics Journal*, 4(1), 1–7.

[50] Kim T.K. (2015). T test as a parametric statistic. *Korean Journal of Anesthesiology*, 68(6), 540.

[51] Kirby K.N. and Gerlanc D. (2013). BootES: An R package for bootstrap confidence intervals on effect sizes. *Behavior Research Methods*, 45(4), 905–927.

[52] Ledoit O. and Wolf M. (2011). Robust performances hypothesis testing with the variance. *Wilmott*, 2011(55), 86–89.

[53] Lim T.S. and Loh W.Y. (1996). A comparison of tests of equality of variances. *Computational Statistics & Data Analysis*, 22(3), 287–301.

[54] Lu M.C. and Chang D.S. (1997). Bootstrap prediction intervals for the Birnbaum-Saunders distribution. *Microelectronics Reliability*, 37(8), 1213–1216.

[55] Martin M.A. (1990). On bootstrap iteration for coverage correction in confidence intervals. *Journal of the American Statistical Association*, 85(412), 1105–1118.

[56] Martini D.D. (2012). Stability criteria for the outcomes of statistical tests to assess drug effectiveness with a single study. *Pharmaceutical Statistics*, 11(4), 273–279.

[57] McNutt M. (2014). Journals unite for reproducibility. *Science*, 346(6210), 679–679.

[58] Miller J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review*, 16(4), 617–640.

[59] Mojirsheibani M. (1998). Iterated bootstrap prediction intervals. *Statistica Sinica*, pp. 489–504.

[60] Mojirsheibani M. and Tibshirani R. (1996). Some results on bootstrap prediction intervals. *Canadian Journal of Statistics*, 24(4), 549–568.

[61] Rizzo M.L. (2008). *Statistical Computing with R*. Boca Raton ; London : CRC Press.

[62] Rosenkranz G.K. (2014). Bootstrap corrections of treatment effect estimates following selection. *Computational Statistics & Data Analysis*, 69, 220–227.

[63] Rubin D.B. (1981). The bayesian bootstrap. *The Annals of Statistics*, 9, 130–134.

[64] Senn S. (2002). A comment on replication, p-values and evidence SN Goodman, Statistics in Medicine 1992; 11: 875-879. *Statistics in Medicine*, 21(16), 2437–2444.

[65] Shao J. and Chow S.C. (2002). Reproducibility probability in clinical trials. *Statistics in Medicine*, 21(12), 1727–1742.

[66] Silverman B. and Young G. (1987). The bootstrap: to smooth or not to smooth? *Biometrika*, 74(3), 469–479.

[67] Welch B.L. (1947). The generalization of 'Student's' problem when several different population varlances are involved. *Biometrika*, 34(1-2), 28–35.

[68] Welch B.L. (1951). On the comparison of several mean values: an alternative approach. *Biometrika*, 38(3-4), 330–336.

[69] Yin T.S. and Othman A.R. (2009). When does the pooled variance t-test fail? *African Journal of Mathematics and Computer Science Research*, 2(4), 56–62.

[70] Zabell S.L. (2008). On Student's 1908 article "the probable error of a mean". *Journal of the American Statistical Association*, 103(481), 1–7.