Contributions to Statistical Reproducibility and Small-Sample Bootstrap

ANDREA SIMKUS

Contributions to Statistical Reproducibility and Small-Sample Bootstrap

Andrea Simkus

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
University of Durham
England

June 2023

Dedicated to

My beloved children, my caring husband, and my inspiring parents

Contributions to Statistical Reproducibility and Small-Sample Bootstrap

Andrea Simkus

Submitted for the degree of Doctor of Philosophy
June 2023

Abstract

This thesis consists of three contributions: an investigation of bootstrap methods for small samples, an overview of reproducibility, and advances on the topic of test reproducibility. These contributions are inspired by statistical practice in preclinical research.

Small samples are a common feature in preclinical research. In this thesis, an extensive simulation study is carried out to explore whether bootstrap methods can perform well with such samples. This study compares four bootstrap methods: nonparametric predictive inference bootstrap, Banks bootstrap, Hutson bootstrap, and Efron bootstrap. The thesis concludes that bootstrap methods can provide a useful estimation and prediction inference for small samples. Some initial recommendations for practitioners are provided.

There are no standardised definitions for reproducibility. This work further contributes to the existing literature by classifying reproducibility definitions from the literature into five types, and providing an overview of reproducibility with a focus on issues related to preclinical research, and on statistical reproducibility and its quantification.

This research explores the variability of statistical methods from the statistical reproducibility perspective. It considers reproducibility as a predictive inference problem. The nonparametric predictive inference (NPI) method, which is focused on the prediction of future observations based on existing data, is applied. In this work, statistical reproducibility is defined as the probability of the event that, if the test was repeated under identical circumstances and with the same sample size, the same test outcome would be reached. This thesis presents contributions to NPI reproducibility for the t-test and the Wilcoxon-Mann Whitney test. As one of the prevailing patterns, a test statistic falling close to the test threshold leads to low reproducibility. In a preclinical test scenario, reproducibility of a final decision involving multiple pairwise comparisons is studied.

Declaration

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

Acknowledgements

I would like to wholeheartedly thank my supervisors Professor Frank P. A. Coolen and Dr Tahani Coolen-Maturi. Their professional expertise, inquisitive minds, and kindness have made my PhD journey significantly more pleasant and productive. Professor Coolen has provided numerous detailed and insightful pieces of feedback on my work. Dr Coolen-Maturi has helped me expand my programming skills, while also becoming an extraordinary role-model and inspiration.

This research work has been carried out in collaboration with AstraZeneca, in particular a team from preclinical research. I would also like to express my gratitude to my AstraZeneca supervisors Dr Claus Bendtsen and Dr Natasha A. Carp, who taught me that it is crucial for research to have an impact, deliver practical value and be communicated well. They have also both provided me with many insights into preclinical research and applied statistics.

I found the tour of the laboratories eye-opening, allowing me to understand the story behind the data, and the importance and meaningfulness of the appropriate statistical analysis in preclinical research. This has been a major motivation underpinning my research.

I would like to extend my thanks to my loving husband and family, who have been very helpful and supportive, both practically and emotionally, as well as my friends and colleagues forming a loving and supportive community in which research can only thrive. Lastly, I would like to thank my children, Madeleine and Michael, for providing me with lots of love and cuddles.

Contents

Chapter 1

Introduction

1.1 Overview

This thesis presents three contributions: it explores the performance of bootstrap methods at making an estimation and prediction inference for small samples, it presents an extensive overview of reproducibility, and it presents some advances on the topic of test reproducibility. Small-sample bootstrap and statistical reproducibility are under-explored, yet of considerable importance to scientific research; and they show potential application in preclinical research.

The first research question explored in this thesis is whether a bootstrap method can provide useful inference for small samples. In preclinical research, the sample sizes are usually small. This is mainly due to cost and animal welfare requirements. The main issue of small samples is that it is hard and sometimes impossible to determine the underlying distribution of the data, and some statistical tests require an assumption of an underlying distribution, e.g. the t-test requires that the distribution is approximately Normal. Chapter 2 provides new insights into the performance when it comes to estimating population characteristics, and making prediction inference for small sample sizes for four bootstrap methods, NPI bootstrap (NPI-B), Banks bootstrap (Banks-B), Hutson bootstrap (Hutson-B) and Efron bootstrap (Efron-B). It focuses on data simulated from Normal, Lognormal, Exponential and Mixed-Normal distributions. This study confirms that the NPI bootstrap method performs well at making prediction inference for small sample sizes, but it also presents further findings regarding the smoothened bootstrap

methods, Banks-B and Hutson-B, and their performance when it comes to the estimation of population characteristics.

Bootstrap methods are often used for building confidence intervals, and for estimation of the bias and standard error of a statistic [71]. Section 2.2 will present a summary of the use of bootstrap methods in pharmaceutical research. One of the applications is bootstrap hypothesis testing. Due to small sample sizes, there are many cases where the Normality assumption is assumed incorrectly and thus the t-test is used incorrectly. One can use a nonparametric counterpart, such as the Wilcoxon Mann-Whitney test. However, nonparametric tests are less powerful than parametric ones. This problem can be overcome by the use of a nonparametric bootstrap method, which does not require the assumption of a particular underlying distribution. Therefore, bootstrap method application in hypothesis testing for small samples is what will be explored in this research.

Reproducibility is a highly discussed topic in pharmaceutical settings and in other research fields. A better understanding of the reproducibility of tests is crucial for preclinical research, as a lack of reproducibility contributes to failure rates in drug discovery and development processes, increasing costs, and decreasing efficiency. Chapter 3 presents a literature review on the topic of reproducibility, summarising several important debates. There is no standardised definition of reproducibility and related terms. Chapter 3 classifies the available definitions from the existing literature into five categories, which we refer to as Type A to Type E. Furthermore, this chapter outlines reasons for low reproducibility and suggests ways to improve the reproducibility presented in the literature on reproducibility and it introduces some reproducibility issues related to preclinical research. The main focus of this chapter is on statistical reproducibility. Various interpretations of and debates in statistical reproducibility are discussed, as well as quantification methods offered in the literature.

This work formulates reproducibility as a predictive inference problem. Statistical reproducibility provides inference on the probability that the same test outcome would be reached, if the test was repeated under identical conditions. The nonparametric predictive inference (NPI) method is used to quantify statistical reproducibility.

NPI statistical reproducibility has been developed by Coolen, Maturi-Coolen, Bin-Himd and Alqifari [5, 31, 50, 52, 53]. The first application of NPI to study reproducibility

was presented by BinHimd and Coolen [31,52], who explored NPI reproducibility for simple nonparametric tests – one-sample sign test, one-sample Wilcoxon signed rank test, two-sample rank sum test and the Wilcoxon Mann-Whitney test (WMT) – and they also developed NPI bootstrap [53], which will be introduced in Chapter 2. Alqifari and Coolen [5, 49] developed NPI reproducibility for tests on population quantiles and for a precedence test. Coolen, Marques and Coolen-Maturi [144, 145] studied reproducibility for likelihood ratio tests.

NPI reproducibility has not yet been presented for the t-test, which is a common test used in preclinical research. This thesis contributes to the literature by presenting NPI reproducibility for the t-test and its application in a real-world scenario. P-values and measures of effect size, such as Cohen's d, play a role in decision making. Thus, as part of the statistical reproducibility topic, this work explores whether there is any relationship between reproducibility and p-values or effect sizes. Reproducibility of a final decision reached through multiple pairwise comparison is also studied. This topic has not yet been explored in the literature.

NPI reproducibility probability is traditionally expressed in lower and upper reproducibility probabilities. However, it is challenging to analytically derive exact lower and upper reproducibility probabilities for large sample sizes or for parametric tests. Thus, this research focuses on the estimation of reproducibility probabilities. Two implementations of NPI are used to quantify statistical reproducibility: NPI bootstrap (NPI-B) and sampling of orderings. The NPI-B method provides a point estimate of reproducibility probabilities while the sampling of orderings method provides estimates of lower and upper reproducibility probability. The main focus is on reproducibility calculated via NPI-B. Sampling of orderings is briefly considered in this thesis for both the WMT and the t-test.

The rest of the chapter is organised as follows: Section 1.2 summarises background information on preclinical research. Section 1.3 outlines two pairwise comparison tests, the t-test and the WMT. Nonparametric Predictive Inference (NPI) is presented in Section 1.4 and NPI for reproducibility is introduced in Section 1.5. Sampling of orderings, a method used to calculate estimates of lower and upper NPI reproducibility, is discussed in Section 1.6. Finally, the outline of this thesis is given in Section 1.7.

1.2 Background to preclinical research

Preclinical research encompasses all research that is done before a particular drug is tested on humans. Preclinical research consists of in vivo studies, where initial studies are carried out on rodents and later studies are carried out on animals more similar to humans, such as pigs and dogs; and of in vitro studies, which are carried out on cells, or organ-on-a-chip, such as bone marrow-on-a-chip. This research has been linked to collaboration with AstraZeneca (AZ). The welfare of animals in animal testing is of crucial importance and a culture of care is nurtured in preclinical research. For example, measures are put in place in order to avoid compassion fatigue, i.e. when clinicians become too hard-hearted. AZ is dedicated to the 3R principles [84]: Replacement, Reduction and Refinement. Replacement means avoiding or replacing the use of animals, reduction means minimising the number of animals used and refinement means minimising animal suffering and improving welfare. This work will further elaborate on these principles in Section 3.5, which addresses some issues regarding the reproducibility of studies that specifically relate to preclinical research. Examples of treatment areas in preclinical research are oncology, asthma and diabetes. There is a variety of different types of studies: for example, in an efficacy study, the disease response to the drug (or different doses of the drug) is examined in an animal model, whereas a toxicity study examines the safety profile of the drug, or different doses of the drug.

In clinical trials, there are regulations that make sure that clinicians follow statistician's advice, which is not the case in preclinical studies. However, at AstraZeneca, all scientists have to follow Good Statistical practice (GSP) [160]. GSP is built on 9 statistical principles: animal numbers, analysis, randomisation, experimental procedures, design, blocking, monitoring, controls and blinding. Also, ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines [10] are followed. In experimental design, sample sizes and the number of groups is decided based on power; simulations are made based on historical data, effect size and variance, where both the worst-case and plausible scenarios are simulated. There is discussion between biologists and statisticians about the experimental design. Statisticians need to have good social skills, as well as good data exploration skills. Statisticians advise biologists and carry out statistical health checks (SHC), they also educate biologists.

The pharmaceutical industry has moved from standardisation to embracing variability because of standardisation fallacy. One of the reasons is the translational failure, i.e. failure at clinical level following a successful preclinical stage. This failure is partly due to the fact that animals sometimes react differently to drugs than humans and preclinical research starts with a healthy animal, inducing a medical condition, and subsequently treating it, whereas human patients already have the disease at the treatment stage. This topic is interlinked with reproducibility and will be further addressed in Section 3.5.4.

1.3 Pairwise comparison tests

This thesis explores reproducibility in relation to two pairwise comparison tests: the t-test and the Wilcoxon Mann-Whitney test (WMT). These are the standard parametric and nonparametric tests for testing a difference in central tendency. Both tests are commonly used in preclinical research statistical analysis [93]. Their test assumptions are the same except that the t-test assumes that the data is Normally distributed. Chapter 4 focuses on reproducibility of the two-sample one-sided tests.

Let $X_1, \ldots, X_{n_x}$ be independent and Normally distributed random variables with mean $\overline{x}$, sample standard deviation s_x and sample size n_x. Let $Y_1, \ldots, Y_{n_y}$ be independent and Normally distributed random variables with mean $\overline{y}$, sample standard deviation s_y and sample size n_y. The t-test compares these two random samples. The t-test tests the null hypothesis H_0: $\overline{x} = \overline{y}$ against H_1: H_1: $\overline{x} > \overline{y}$ (in the upper-sided t-test), H_2: $\overline{x} < \overline{y}$ (in the lower-sided t-test) or H_3: $\overline{x} \neq \overline{y}$ (in the two-sided t-test).

The test statistic of the t-test, t, is a standardised value. H_1 is rejected if $t > t_{n+m-2,\alpha}$, H_2 is rejected if $-t < t_{n+m-2,\alpha}$ and H_3 is rejected if $|t| > t_{n+m-2,\frac{\alpha}{2}}$. The calculation of the t-statistic and the number of degrees of freedom (df) depends on whether equal variance of samples is assumed or not. For the equal variance t-test, the t-statistic is calculated using Equation (1.1).

$$t = \frac{\overline{x} - \overline{y}}{s_p \sqrt{\frac{1}{n_x} + \frac{1}{n_y}}}, \text{ where } s_p = \sqrt{\frac{(n_x - 1)s_x^2 + (n_y - 1)s_y^2}{n_x + n_y - 2}} \tag{1.1}$$

and $df = n_x + n_y - 2$. The calculated s_p is the pooled standard deviation.

The unequal variance t-test is called the Welch t-test and its t-statistic is calculated via Equation (1.2).

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}} \tag{1.2}$$

and the degrees of freedom are calculated via Satterthwaite's correction displayed in Equation (1.3).

$$df = \frac{(n_x - 1)(n_y - 1)}{(n_x - 1)c_2^2 + (n_y - 1)c_1^2}, \text{ where } c_1 = \frac{\frac{s_x^2}{n_x}}{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \text{ and } c_2 = 1 - c_1 \tag{1.3}$$

As a complementary to the t-test, Cohen's d is an often used measure of the standardised effect size for comparisons of two samples. Throughout this thesis, the term Cohen's d is used rather than the term standardised effect size to mimic the terminology used in preclinical research. Cohen's d is given by Equation (1.4) [43], where s is the average of the two individual sample standard deviations, s_x and s_y, i.e. $s = \frac{s_x + s_y}{2}$. Here s is used instead of s_p because two simulated samples in pairwise tests in Chapter 4 are always of the same size, and the samples in the preclinical scenario in Chapter 4 are nearly of the same size while their standard deviations are similar.

$$d = \frac{(\overline{x} - \overline{y})}{s} \tag{1.4}$$

For the Wilcoxon Mann-Whitney test (WMT), there are $N = n_x + n_y$ observations of $X_1, \ldots, X_{n_x}$ and $Y_1, \ldots, Y_{n_y}$, which are two independent and identically distributed random samples. These samples are mutually independent and their distributions are continuous. $X_1, \ldots, X_{n_x}$ is a sample from some distribution F and $Y_1, \ldots, Y_{n_y}$ is a sample from some distribution G. WMT tests the null hypothesis H_0: $F(t) = G(t)$ for every t, i.e. X and Y variables have the same probability distribution but the common distribution is not specified, against the alternative hypothesis H_1: $G(t) = F(t - \delta)$ for every t. To test whether the distribution F is shifted to the left of G, i.e. for positive δ, the upper-sided WMT is used, whereas to test whether the distribution F is shifted to the right of G, i.e. for negative δ, the lower-sided WMT is used. To test whether there is any shift, the double-sided WMT is used.

Nonparametric WMT ranks the observations from the combined two samples of $N = n_x + n_y$ X-values and Y-values. The test statistic of the WMT is the sum of the ranks for observations from the Y sample. Let S_1 denote the rank of $Y_1, \ldots, S_{n_y}$ denote the rank of Y_{n_y}. These ranks are within the combined sample. WMT leads to the statistic Z

$$Z = \sum_{j=1}^{n_y} S_j. \tag{1.5}$$

For the upper-tailed two-sample WMT, H_0 is rejected if $Z \geq Z_\alpha$. For the lower-tailed one-sided two-sample WMT, H_0 is rejected if $Z \leq n_y(n_x + n_y + 1) - Z_\alpha$. Z_α is the critical value and it can be read from the tables, which can be found in [108].

For large samples of X and Y (i.e. with $n_x > 10$ or $n_y > 10$), large sample approximation is used. This approximation is based on the suitably standardised asymptotic Normality of Z [108]. Under the H_0, the mean and the variance of Z are:

$$E_0(Z) = \frac{n_y(n_x + n_y + 1)}{2} \tag{1.6}$$

$$var_0(Z) = \frac{n_x n_y(n_x + n_y + 1)}{12} \tag{1.7}$$

and the approximate Z is denoted by Z_*.

$$Z_* = \frac{Z - E_0(Z)}{\{var_0(Z)\}^{\frac{1}{2}}} = \frac{Z - \frac{n_y(n_x+n_y+1)}{2}}{\{\frac{n_x n_y(n_x+n_y+1)}{12}\}^{\frac{1}{2}}} \tag{1.8}$$

For the upper-tailed one-sided test, H_0 is rejected if $Z_* \geq z_\alpha$ and for the lower-tailed one-sided test, H_0 is rejected if $Z_* \leq -z_\alpha$. For $\alpha = 0.05$, $z_{0.05} = 1.645$. The large sample approximation will be used in the analysis of datasets in Section 4.7.

1.4 Nonparametric predictive inference (NPI)

Nonparametric predictive inference (NPI) [48, 51] has been applied in many areas, for example, in finance [15], system reliability [54], operations research [47] and receiver operating characteristic analysis [58]. NPI is based on Hill's assumption $A_{(n)}$, which is a post-data assumption that gives conditional probabilities for a future observation [106].

Let $X_1, \ldots, X_n, X_{n+1}$ be real-valued exchangeable random quantities. $X_1, \ldots, X_n$ are observed and the aim is to make inference based on future observations via Hill's assumption. The ordered observed values are $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ and let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$, or use known or assumed bounds for the support of the random quantities, say $x_{(0)} = L$ and $x_{(n+1)} = R$ [47]. Then for the future observation X_{n+1}, based on n observations, the assumption $A_{(n)}$ is [47]:

$$P(X_{n+1} \in (x_{(j-1)}, x_{(j)})) = \frac{1}{n+1}, \text{ for } j = 1, 2, \ldots, n+1. \tag{1.9}$$

This means that X_{n+1} is equally likely to be in any of the intervals created by the ordered observed data. Note that under $A_{(n)}$ it is assumed that there are no ties. In the NPI framework, ties can be dealt with by breaking them by a very small amount [56, 57, 146]. When relevant, `jitter` function in R is used in this research.

The NPI approach can also be used for multiple future observations via the consecutive application of Hill's assumption $A_{(n)}, A_{(n+1)}, \ldots, A_{(n+m-1)}$, which together are denoted by $A_{(\cdot)}$ [49]. An ordering O_i represents the possible positions of the $m > 1$ future observations relative to the n data observations. There are $\binom{n+m}{n}$ possible orderings of m among the n observations, and under $A_{(\cdot)}$ all these orderings are equally likely [49, 88], as is implied by Equation (1.10). Let S_j^i denote the number of future observations in the interval $I_j = (x_{(j-1)}, x_{(j)})$ given the specific ordering O_i, where $i = 1, \ldots, \binom{n+m}{n}$ and $j = 1, \ldots, n+1$. Here s_j^i are non-negative integers and $\sum_{j=1}^{n+1} s_j^i = m$.

$$P(\bigcap_{j=1}^{n+1} \{S_j^i = s_j^i\}) = P(O_i) = \binom{n+m}{n}^{-1}, \; i = 1, \ldots, \binom{n+m}{n} \tag{1.10}$$

Any specific ordering only specifies the number of future observations in each interval I_j, no assumptions are made about the exact location of the future observations within the interval I_j.

Uncertainty is traditionally expressed using lower and upper probabilities in the NPI framework [13]. This is because the exact position of future points is not relevant within this framework. What matters is that a future point belongs to an interval I_j between two consecutive observations $x_{(j-1)}$ and $x_{(j)}$. Lower probability of the event of interest A is the maximum lower bound for the precise probability of the event A. In hypothesis testing, the event A could be either the rejection or non-rejection of the null hypothesis.

Informally, lower reproducibility probability reflects the evidence certainly in favour of the event A [13]. In the NPI framework, lower probability takes into account only the orderings of m future observations among the n current observations for which the event A has to hold [49]. Upper probability is the minimum upper bound for the event A, reflecting all evidence that could be possibly in favour of the event A. In the NPI framework, upper probability takes into account all orderings for which the event A could hold [49].

A note on exchangeability

Hill's assumption requires that random quantities are exchangeable. Exchangeability does not imply a form of dependence, else one could not learn from the observations about non-observed random quantities, but one cannot just add any form of dependence. For example, one could not add a constraint, nor another known form of dependence. Exchangeability also does not imply a form of independence. If random quantities X and Y are independent, then any information we get (or assume) about X does not change our knowledge or beliefs about Y. $A_{(n)}$ is employed in cases where there is little knowledge about the random quantity of interest or when a choice has been made not to use this information, thus independence is not a suitable assumption.

In the NPI framework, exchangeability implies - for real-valued quantities - that the orderings are equally likely before observing the values. In a frequentist statistics setting, $A_{(n)}$ then fills in the values of n observations and hence leads to the $\frac{1}{n+1}$ probability for the future observation to be inside each interval between two consecutive observations. Strictly speaking, if one would attempt a minimal formulation, $X_1, \ldots, X_{n+n}$ (for n future observations) would not need to be exchangeable, as only the $A_{(.)}$ assumptions are needed. Hence the exchangeability of the first n (which are being observed) would not be required. Nevertheless, it makes sense to assume that all the random quantities are exchangeable.

1.5 NPI for reproducibility probability

Reproducibility and statistical reproducibility are widely discussed topics and their definitions are not clearly defined in the existing literature, as will be demonstrated in Chapter 3. This thesis focuses on statistical reproducibility and it interprets it as a pre-

diction problem. It narrows statistical reproducibility down to the variability of statistical methods, which exists due to the variability of data, rather than further aspects of reproducibility. This thesis adopts the following definition of statistical reproducibility: the probability of the event that, if a test was repeated under identical circumstances and with the same sample size, the same test outcome would be reached. The classical frequentist approach is unsuitable for solving a predictive problem, thus NPI, as described in Section 1.4, is employed. NPI is focused on future observations, making it a good approach for inference on reproducibility.

In the setting of hypothesis tests, the test outcome means the rejection or non-rejection of the H_0. Statistical reproducibility can be determined in the following manner: After performing a hypothesis test on the original sample of size n, we determine whether or not to reject the H_0 based on the value of the test statistic. We then predict a future sample of size n, where all orderings of the n future observations among the n actual data observations are equally likely. Next, we determine whether H_0 is certainly rejected, possibly rejected, possibly not rejected, or certainly rejected for each ordering of the future observations. For the lower reproducibility probability, we count all orderings for which the conclusion is certainly the same as for the actual test for the lower reproducibility probability. For the upper reproducibility probability, we include the 'possibly' orderings where the conclusion is the same as for the actual test.

The NPI reproducibility probability does not imply anything about getting the test outcome 'right'; for that, traditional aspects of hypothesis testing, such as the level of significance, power and other related post-data metrics, are relevant. To calculate NPI reproducibility, the full data set is required; different data with the same value of the test statistics can lead to different reproducibility values. To provide high quality analysis, other statistical methods including power, effect size (ES) and p-value could be used together with statistical reproducibility. NPI reproducibility is another property that allows decision makers to extend the decision-making capacity in making more robust decisions. In pharmaceutical research, at each stage of the process one needs to decide whether to continue to a further study or repeat the test. NPI reproducibility probability is another metric to help the team of practitioners and statisticians make this decision.

As will be explained in Section 3.5, in preclinical research, the discussion of repro-

ducibility is focused on adhering to good statistical practice and on embracing the inevitable variability caused by the use of animals. NPI reproducibility research is not concerned with the issue of deviations stemming from the fact that animal testing is never carried out under identical conditions, e.g. mice have slightly different properties and new experiments are often carried out in different laboratories. This research is solely and exclusively limited to the investigation of reproducibility of statistical tests, based on the original test scenario data and the description of the data and the statistical analysis. While much of the discussion on reproducibility in the literature is on whether an experiment can actually be reproduced under similar circumstances, that is irrelevant for NPI reproducibility because NPI reproducibility does not include an actual second experiment.

1.6 Sampling of orderings to estimate NPI-RP

NPI reproducibility is customarily expressed in lower and upper reproducibility probabilities, $\underline{RP}$ and $\overline{RP}$, because $A_{(n)}$ is not a sufficient assumption to calculate precise probability of an event A, as discussed in Section 1.4. BinHimd [31] presented a method for calculating $\underline{RP}$ and $\overline{RP}$ for the upper-tailed two-sample Wilcoxon Mann-Whitney test (WMT) by considering all the orderings of m future observations among the n current observations. This method will be further described in Section 4.5.1. However, calculating precise $\underline{RP}$ and $\overline{RP}$ is computationally not feasible for larger sample sizes, as shown in BinHimd's thesis [31], but it is possible to estimate them. NPI bootstrap (NPI-B), which will be introduced in Section 2.3.3, is one of the tools that can be used to estimate reproducibility probability. However, NPI-B provides a point estimate of reproducibility probability, it does not enable presentation of the results in terms of imprecise reproducibility probabilities. This thesis explores a method that calculates estimates for lower and upper reproducibility probabilities through sampling of orderings. Instead of analytically deriving lower and upper reproducibility through considering all the possible orderings, only a selected amount of orderings is sampled. Sampling of orderings for the likelihood test was presented by Marques et al. [55, 144, 145].

Sampling of orderings reduces the computing time and enables the calculation of estimates for imprecise reproducibility probabilities for larger number of original points. This

method will be illustrated in Chapter 4: the sampling of orderings method is employed to calculate the estimates for NPI-RP for the Wilcoxon Mann-Whitney test in Section 4.5 and then the heuristics for approximating NPI-RP lower and upper reproducibility probability for the t-test are presented in Section 4.6.

1.7 Outline of the thesis

The rest of this thesis is organised as follows: Chapter 2 compares the performance of different bootstrap methods for estimation and prediction inferences via a simulation study. The focus is on making inferences for small samples. Both Normal and other distributions are considered in this simulation study. Chapter 3 introduces the topic of reproducibility. It addresses the issue that there are no standardised definitions for reproducibility and it classifies the definitions from the literature into five types. Reasons for bad reproducibility and suggestions for improvement offered in the literature are discussed. The chapter further provides insights into the discussions regarding statistical reproducibility. An overview of metrics for quantification of reproducibility is presented and, finally, the NPI method for quantification of reproducibility is placed in the context of the wider literature. Chapter 4 explores reproducibility for pairwise tests. An algorithm for calculating bootstrapped reproducibility for the pairwise t-test is presented and explored in a simulation study; and an algorithm for calculating the reproducibility of the final decision, when multiple pairwise comparisons are carried out, is introduced. Both algorithms are applied to a real-life scenario from preclinical research. This part of the chapter is based on a paper co-published by the author of this thesis in the *Statistical Methods in Medical Research* journal [192]. Estimation of lower and upper reproducibility probabilities via the sampling of orderings is illustrated on the WMT and heuristics for approximating NPI-RP lower and upper reproducibility probability for the t-test are discussed. NPI reproducibility estimation via both the NPI-B method and the sampling of orderings is further illustrated on the rate of growth data for not Normally distributed datasets. The thesis concludes with a summary of the findings and with the formulation of future research questions in Chapter 5. Calculations have been done using R versions 3.2.4 (Chapter 4) and 3.6.3 (Chapter 2). R code is provided in Appendix C.

Chapter 2

Bootstrap performance for small samples

2.1 Introduction

Smaller samples are common in biomedical research [177]. These smaller samples are limited in the ability to justify model assumptions underlying most classic statistical techniques. Within the setting of biomedical research, there are often few historical studies to guide the decision-maker on assumptions about the underlying distribution of the data. Thus, there is an increased possibility of making decisions based on wrong assumptions. For example, sometimes Normal distribution is assumed incorrectly with small sample sizes. In such cases, the data might contain outliers, or the Normal distribution might not fit the data well. Therefore, there is practical value in exploring bootstrap methods for smaller samples.

In the literature, we have not encountered Efron-B as a commonly used method in preclinical studies with small sample sizes. This could be because the most commonly used bootstrap method for the quantification of uncertainty in the estimation of parameters, Efron-B, is not considered to be a reliable method for very small sample sizes [42, 201]. Efron-B has primarily been created for large samples, where it shows good performance in estimation. The main argument supporting Efron-B, the asymptotic argument, is based on the fact that the empirical distribution, from which a sample is taken in Efron-B, converges to the real underlying population distribution if the number of data increases

to infinity [27]. However, Efron-B does not perform well in the case of estimation for small samples as it does not provide good coverage [18]. Efron [78] proposed bias-corrections to his bootstrap method, bias-corrected and accelerated bootstrap (BC_a) and approximate bootstrap confidence (ABC) intervals, to improve the bootstrap coverage, but he admits that "their coverage accuracy can still be erratic for small sample sizes" [78].

The purpose of this chapter is to compare four bootstrap methods: Efron bootstrap (Efron-B), Banks bootstrap (Banks-B), NPI bootstrap (NPI-B) and Hutson bootstrap (Hutson-B), when applied with small samples. This chapter provides new insights into the performance in making estimation and prediction inferences for small sample sizes for these four bootstrap methods. The investigation is done via a simulation study carried out for data simulated from Normal, Lognormal, Exponential and Mixed-Normal distributions. This chapter considers the estimation of various population characteristics: mean, median, variance, first quartile (Q1), third quartile (Q3) and interquartile range (IQR). The main research question is whether a bootstrap method can provide useful information when used with small samples. Another aim of the study is to provide some initial recommendations on the small-sample bootstrap for practitioners.

Banks-B is not a well-known bootstrap method, but it has potential to successfully quantify the uncertainty in sample-based estimates of population characteristics for small samples [18]. Hutson-B [111] has been introduced as a new quantile function estimation method for generating bootstrap samples, rather than a bootstrap method. This thesis views it, however, as a bootstrap method. NPI bootstrap has been developed for prediction inference. Efron-B has been used in bootstrap hypothesis testing to calculate approximate p-values when comparing means of two samples. Initial investigations have been done for small sample sizes by Dwivedi et al. [73]. Section 2.6 extends this study to include Banks-B and NPI-B.

This chapter begins by reviewing some of the uses of bootstrap methods in pharmaceutical research in Section 2.2. These are mostly based on Efron bootstrap. This section focuses on application of bootstrap methods in both preclinical and clinical research. Most of the literature has focused on large samples and clinical studies rather than preclinical studies with small samples. Section 2.3 introduces four bootstrap methods (Efron-B, Banks-B, NPI-B, Hutson-B) and percentile and BC_a confidence intervals.

Section 2.4 assesses the performance in the estimation of population characteristics for the four bootstrap methods. This is followed by a comparison study of the bootstrap methods' performance in prediction in Section 2.5. Section 2.6 explores bootstrap hypothesis testing, including Banks-B and NPI-B in the investigation. Section 2.7 summarises findings of this chapter and outlines suggestions for potential future research.

2.2 Bootstrap methods in pharmaceutical research

This section will introduce, rather than give a completely overview of, some of the practical uses of the bootstrap method. It focuses on the published literature utilising bootstrap methods in pharmaceutical research. Most of these examples are from clinical research and they focus on large sample sizes. These applications will be introduced first. We encountered only few studies focusing on small sample sizes and only one of them presented a real life example from preclinical research. The possible explanation for bootstrap methods not being commonly used with small samples is that the most commonly known and applied bootstrap method is Efron's bootstrap, which has not shown good performance for small samples, as discussed in Section 2.1. This overview serves as both a motivation for the exploration carried out in this chapter, and as a list of possible areas that could be explored with bootstrap methods for small samples. This chapter will show that the bootstrap method has a potential use for small samples. Bootstrap applications for large, medium and small samples are presented in Sections 2.2.1, 2.2.2 and 2.2.3, respectively. There is no universally accepted definition of small, medium and large sample size and it is outside the scope of this thesis to discuss what sample size is still small and what is not. This thesis will focus on sample sizes $n = 4, 6, 8, 10$ when assessing the bootstrap method's performance in estimation and on $n = 4, 6, 8, 10, 20$ when assessing the bootstrap performance in prediction.

2.2.1 Applications for large samples

In practice, the Efron's bootstrap method is often used to estimate bias and standard error, or to construct confidence intervals [71]. An accurate estimate of the uncertainty associated with parameter estimates is important to avoid misleading inferences. Walters

and Campbell [204] explored Efron-B use in the estimation of standard errors and BC_a confidence intervals, which will be defined in Section 2.3.5, for parameters calculated during the analysis of health-related quality of life outcomes (HRQoL). An example of such parameter estimate, $\hat{\theta}$, is the mean difference (intervention mean - control mean). HRQoL data are often recorded on an ordinal scale and they typically have bounded, discrete and skewed underlying distributions and the sample sizes are large. In four studies, the bootstrap method was compared to the conventional statistical methods, such as the linear regression. Conventional ordinary least squares estimates of standard error and confidence interval for the group regression coefficient were compared with their bootstrap counterparts. The sample sizes in these studies were large: they range from 100-250. Walters and Campbell [204] concluded that the conventional statistical methods and the bootstrap methods produced similar results, i.e. similar standard errors and confidence intervals.

Efron-B has also been used in hypothesis testing [97]. Walters and Campbell [204] explored the use of Efron-B for hypothesis testing, using algorithm from Efron and Tibshirani [78, p.224], when analysing the above mentioned four studies focused on HRQoL. Walters and Campbell [204] concluded that the bootstrap method led to similar p-values as the conventional statistical methods. The explanation for this conclusion could be linked to the use of sufficiently large sample sizes. According to the central limit theorem, sample means are approximately Normally distributed for large sample sizes. Thus, the t-test is an appropriate test for large samples. Similarly, Efron-B is suitable for large samples, as explained in Section 2.1.

Bootstrap methods have also been applied in power and sample size calculations [164,170]. Traditional methods for power and sample size calculations require an estimate of treatment effect and sample variance, and they are based on known distributions [164]. However, there are cases where traditional power calculation methods cannot be used, i.e. when the Normal distribution cannot be reasonably assumed, or where the study decision is based on co-primary outcome measurements [164]. Co-primary outcome measurements were not defined in [164] but, in a different source, co-primary endpoints were defined as "two or more trial endpoints, each measured among a group of patients and each equally important in determining efficacy" [39], where the endpoint is a clinical variable reflecting

the condition of a disease. Peng et al. [164] illustrated the application of the bootstrap method in the power analysis and sample size estimation on two examples of clinical trial designs, both focused on large samples. In the first example, in a study of a drug for the Alzheimer's disease, two co-primary outcome variables, one with categorical data and one with data on a continuous scale, were compared between treatment and placebo group. The focus was on sample sizes varying from 75 to 125. For each variable, p-value was calculated: p_1 and p_2. H_0 was not rejected if $p_1 \geq \alpha$ or $p_2 \geq \alpha$ for some given α. The aim of the study was to determine the sample size which would provide 80% dual outcome power. Dual outcome power is "the probability of observing a significant drug versus placebo comparison with respect to both primary efficacy variables" [164]. Data for each outcome variable of each group were generated. For the continuous variable, random data were generated, based on the given population mean and standard deviation. For the categorical variable, a probability distribution of a categorical variable was used to calculate its cumulative distribution, and pseudo-random numbers were generated and mapped "to categorical values by associating the quantile to that random number" [164]. In the second example, bootstrap power analysis was carried out for the stratified Wilcoxon test for sample sizes varying from 100 to 900. Power was calculated in two steps: First, a large number of bootstrap sample data sets were generated from the original trial data set. Secondly, the original statistical test was applied to each bootstrap data set generated and power was estimated as the percentage of the times the null hypothesis was rejected for these bootstrap samples [164].

2.2.2 Applications for medium samples

So far, the cited literature focused mostly on large samples. The focus of this section will be on medium sample sizes. Barber and Thompson [20] compared conventional methods, such as the t-test and the Wilcoxon Mann-Whitney test, and bootstrap hypothesis testing for comparison of the arithmetic mean of costs in two treatment groups. The cost data are from health economic evaluations, which guide health care policy decisions. The data are often highly skewed [20]. Two examples were given, one of large sample size ($n_1 = 70$ and $n_2 = 74$), and second of medium sample size ($n_1 = 18$ and $n_2 = 14$). It could be argued that in the first example the sample size is medium, not large and

that in second example the sample size is small, not medium. Barber and Thompson concluded that for the large sample size study, the t-test yielded similar results as the bootstrap hypothesis test. For small to medium sample sizes, Barber and Thompson [20] recommended to report bootstrap analysis results, or to use these to check robustness of parametric methods. The named advantage of using bootstrap methods was the avoidance of having to make assumptions about the underlying distribution of the data. However, Barber and Thompson [20] highlighted that bootstrap methods rely on the assumption that the empirical distribution adequately represents the true distribution of the data. Barber and Thompson [20] recommended using BC_a or bootstrap-t confidence intervals rather than percentile confidence intervals because coverage error for percentile confidence intervals can be large if the distribution of $\hat{\theta}$ is not symmetrical around the observed value [20]. The applicability of bootstrap-t confidence intervals for the estimation of location statistics has also been suggested by Efron and Tibshirani [78].

The bootstrap method has been used in human immunophenotyping research. Holmes and He [109] employed Hutson-B, which will be introduced in Section 2.3.4, and they referred to it as $\mathscr{Q}(n)$-bootstrap. Clinical studies in this area usually have small and wide datasets, $1 < n < 50$ of human participants, and, for each participant, many parameters $1 < p < 1000$ are estimated. The underlying distribution of the data cannot be ascertained. Thus, the bootstrap method was chosen to estimate immune parameter. $\mathscr{Q}(n)$-bootstrap presumes linearity for extrapolation but it does not assume any stronger assumption [109]. Holmes and He [109] gave an example of medium size participants of study, with $n = 35$ participants, related to seasonal dose of influenza vaccine. The study was interested in age-related changes in immune features and simple linear regression was applied. The sample order statistics of the regression residuals were used for the estimation of the quantile function [109]. Resampling with $B = 2500$ bootstrap samples was carried out from residuals and the stratified sampling study design was used. The goal of the study was to calculate confidence intervals on parameter estimates, imposing minimum assumptions on the data, while obtaining accurate and not too narrow confidence intervals in order to achieve minimal bias, low variance and interpretability [109]. Holmes and He [109] concluded that confidence intervals based on Efron-B with percentile CI were narrower than for $\mathscr{Q}(n)$-bootstrap, Hutson-B, for approximately 68% of the 229 features.

$\mathscr{Q}(n)$-bootstrap had better coverage probabilities compared to Efron-B with both percentile CI and bootstrap-t CI, and compared to smoothed kernel quantile estimator [186]. Holmes and He [109] credited the improvement of the coverage probabilities to the tail extrapolation.

2.2.3 Applications for small samples

Tsukamoto et al. [200] used the bootstrap method in the diagnostics of neurodegenerative diseases. When evaluating images from positron emission tomography (PET) or single-photon emission computed tomography (SPECT), it can be assumed that data distribution may be inappropriate and, thus, it is appropriate to choose a method which does not require an assumption about the underlying distribution of the data. Tsukamoto et al. [200] proposed the use of nonparametric bootstrap, Efron-B, or smoothed bootstrap, a slightly adjusted Hutson-B, in the statistical evaluation of the decrease of regional cerebral blood flow (rCBF), a measure of local neuronal activity, in a SPECT image. The two bootstrap methods were used to calculate a standardised distribution of the Z-score. Z-scores were calculated for each pixel value at (k, l) coordinates of an image for both the control and patient data set of sample size n. Pixels are the smallest components in the digital image. The decrease of rCBF at the pixel was considered statistically significant when a Z-score exceeded a threshold, T_0 [200]. In the example given, control data set of $n = 95$ images was studied. In the simulation, subsamples from the original dataset of small to medium sample sizes, $n = 5, 10, 15, 20, 30, 40$, were considered. For each sample size, 20 datasets were subsampled. For each sample, $B = 5000$ bootstrap samples were created. Tsukamoto et al. [200] concluded that both bootstrap methods produced more consistent results than traditional methods for small samples, $n = 5, 10$, and in cases where the control set was small, the smooth bootstrap method, Hutson-B, was recommended.

Dwivedi et al. [73] performed an extensive simulation study to compare the nonparametric bootstrap test with standard parametric, nonparametric, and permutation tests, for comparisons of means of two independent samples, two dependent samples, and more than two independent samples, for data of various sample sizes, $n = 3, 4, 5, 6, 7, 8, 9, 10, 15$, from both Normal and skewed underlying distribution. As discussed in Section 2.1, min-

imum or even no assumptions can be made about the underlying distribution of the data with small samples, and Efron-B does not provide good coverage for small samples [18]. Dwivedi et al. [73] addressed the latter issue by drawing Efron-B samples from the combined original sample. Note that Dwivedi et al. used the phrase *pooled sample*, the phrase *combined sample* will be used in this chapter instead, as the term *pooled* is usually associated with variance. Hall and Wilson [97] also advised that resampling should reflect the null hypothesis, as this increases the power of the bootstrap test. Let $\hat{\theta}$ be the estimate of the data characteristics calculated from the original data sample and $\hat{\theta}^*$ be the estimate calculated from the bootstrap sample. Hall and Wilson [97] explained that resampling $|\hat{\theta}^* - \hat{\theta}|$ is more meaningful than resampling $|\hat{\theta}^* - \theta_0|$. This is because if θ_0 is far from the true value of θ, then the difference $|\hat{\theta}^* - \theta_0|$ will not appear large compared to the nonparametric bootstrap distribution of $|\hat{\theta}^* - \theta_0|$ and the bootstrap test is less likely to reject the null hypothesis even in cases when the alternative hypothesis is true. Moreover, drawing bootstrap samples from uncombined original samples leads to less resampling variability [73]. Dwivedi et al. [73] concluded that the pooled nonparametric bootstrap t-test is preferable to other statistical methods for small sample size studies for the comparison of two means. This is especially the case when comparing two datasets with unequal variances, unequal sample sizes, and with underlying distributions, which are not Normal [73].

Apart from the simulation study, Dwivedi et al. [73] presented two examples from clinical studies. The first example was a clinical study on epilepsy, which compared the percent seizure reduction between an active arm with sample size $n_{\text{active}} = 6$ and a control arm with sample size $n_{\text{control}} = 5$, among subjects who had more than 18 seizures per month, using unpaired Student's t-test, Welch t-test, nonparametric bootstrap t-test, Wilcoxon rank sum test, and asymptotic permutation t-test [73]. Wilcoxon rank sum test did not reject H_0, whereas the other tests did. Moreover, for each treatment group where $n_{\text{control}} = 10$ and $n_{\text{active}} = 7$, the change in seizure frequency from baseline to post intervention among subjects who had more than 14 seizures per month was compared using paired tests. For the active group, all paired tests found there was reduction in seizure frequency.

The second example was a clinical trial on motivational interviewing which sought to

improve treatment engagement. Two groups, motivational intervention (MI) and standard intervention (SI) were compared, using unpaired Student's t-test, Welch t-test, nonparametric bootstrap t-test, Wilcoxon rank sum test, and asymptotic permutation t-test. The treatment retention and substance use at 28 days and 84 days after randomisation were recorded for both groups. Data from each treatment group were randomly selected, with different sample sizes. Variety of samples sizes were explored, e.g. $n_{\text{MI}} = 6, n_{\text{SI}} = 3$; $n_{\text{MI}} = 20, n_{\text{SI}} = 10$; and $n_{\text{MI}} = 173, n_{\text{SI}} = 177$. The conclusion made was that for unequal sample sizes and unequal variances, p-values obtained using the Welsch t-test and the nonparametric bootstrap t-test were similar but different from p-values obtained through other tests. Bootstrap hypothesis testing will be further explored in conjunction with Banks-B and NPI-B in Section 2.6.

Lastly, this section briefly outlines an article that focused both on preclinical study and small sample size. Mager and Göller [141] discussed the use of bootstrap methods for data that does not follow the Normal distribution in safety assessment in preclinical pharmacokinetics and in toxicokinetics. The sample sizes in safety assessment are usually small. In such studies, concentrations are recorded at multiple time points and the statistics of interest are the standard error of $\text{AUC}|_0^{t_K}$ and the arithmetic mean of $\text{AUC}|_0^{t_K}$. Here AUC represents the area under the concentration-time profile. For a specific dosage, the concentration-time profile plots the exposure to drug versus time after the dosage. Let t_K be a time point. $\text{AUC}|_0^{t_K}$ measures exposure to a drug from time point 0 to time point t_k; it is the area underneath the curve between time point 0 and time point t_k. Two bootstrap methods, pseudoprofile-based bootstrap and the pooled data bootstrap, were employed. Introduction of these two bootstrap methods is beyond the scope of this thesis. Data from three different pharmacokinetic models were analysed. Sample sizes in these models were small: $n = 4$ or $n = 5$. The two bootstrap methods showed to be powerful tools in the safety assessment in the cases where data were not Normally distributed or when it was required to estimate additional secondary pharmacokinetic parameters and their variability. The named advantage of using the bootstrap method, as opposed to the standard method, was that the bootstrap method did not require assumptions about the underlying distribution and both the secondary pharmacokinetic parameters and their variability (such as standard deviations and standard errors) could be assessed [141].

2.3 Bootstrap methods

Bootstrap [78] is one of the resampling methods available to a practitioner, together with the jacknife method [185], the delta method [42], subsampling [42], permutation tests [71, 78], randomisation tests [59], the cross-validation method [78], and Monte Carlo methods [59]. Resampling methods allow the computation of a variety of statistics from limited data while making minimal distribution assumptions.

The advantages of the bootstrap method compared to other resampling methods are that the conceptual understanding behind and implementation of the bootstrap method is simple and straightforward, and there is a variety of bootstrap methods to choose from which allows for flexibility. The variety of bootstrap methods can also be considered a disadvantage as it makes it harder for practitioners to choose a particular bootstrap method. Other disadvantages are: the bootstrap method requires more computer time than other resampling methods, such as the jacknife method, and the most commonly known bootstrap method, Efron-B, does not show good performance for small samples compared to jacknife which shows a better performance [76]. However, this chapter shows that even the bootstrap method can provide useful inference with small samples. Moreover, bootstrap methods do not require an assumption of a particular distribution.

Bootstrap methods can be divided based on how the population is approximated [71] into nonparametric, semi-parametric [40] and parametric bootstrap. In nonparametric bootstrap, no particular distribution is assumed, whereas in the parametric bootstrap [40] a distribution is assumed, the parameters of which are estimated and these estimates are used to draw bootstrap samples. This thesis will focus on nonparametric bootstrap methods. The consideration of parametric bootstrap is outside the scope of this chapter's simulation study because the focus is on small samples, for which it is usually not possible to accurately determine the distribution.

The main focus of this chapter is on four bootstrap methods: the nonparametric ordinary bootstrap, Efron-B [69, 74, 75, 77, 78], which is a bootstrap method commonly used for the quantification of uncertainty in the estimate, mainly for large sample sizes; Banks-B [18], which is not a well-known bootstrap method but it has a potential to perform well in the estimation of population characteristics for small sample sizes [18]; Hutson-B [111], a bootstrap method utilising semi-parametric quantile function estimator,

which has been applied in practical applications for small samples (see Section 2.2); and NPI-B [53], which has been developed for prediction rather then estimation, nevertheless, it has been also used for estimation in [2].

Banks-B and Hutson-B are two different smoothed bootstrap methods. Smoothened bootstrap methods provide more variability of the bootstrap sample than ordinary bootstrap. Smoothened bootstrap methods overcome a problem of Efron-B for small samples: Efron-B samples underestimate the true variability of the data as there are only a few values to sample from [104].

Initial study into the performance in the estimation of population characteristics of another bootstrap method, the smoothed bootstrap using Gaussian kernel (Kernel-B), has been carried out. Kernel-B is not included in the main study because it involves more issues, including the determination of the smoothing parameter, which would divert the focus of the main investigation. The initial findings for Kernel-B are reported in Appendix A.5.2.

One bootstrap method can be used to create different types of confidence intervals, e.g. the percentile, basic, accelerated, studentised, or bias-corrected, accelerated (BC_a) bootstrap, the test-inversion bootstrap method and the Studentised test-inversion bootstrap method. This chapter focuses on percentile confidence intervals because they are simple to implement and they have been employed by Banks [18] in his simulation study. This thesis also briefly considers BC_a confidence intervals. Efron proposed using BC_a confidence intervals in order to improve the bootstrap coverage [78]. BC_a confidence intervals will be explored in relation to the estimation of population characteristics for small samples. There are many available R packages in CRAN for bootstrap methods, to name some, package `bootstrap` is based on Efron and Tibshirani [78], `boot` is based on Davison and Hinkley [64] and `bcaboot` focuses on calculating bias corrected bootstrap confidence intervals. Further overviews of the bootstrap method can be found in [41, 42, 64, 78, 96, 137, 142, 185].

2.3.1 Efron bootstrap

The Efron's bootstrap method (Efron-B) has also been referred to as nonparametric bootstrap [73]. In this method, the unknown underlying distribution F is replaced with the

empirical distribution F_n of the observed data $x_1, x_2, \ldots, x_n$ [78]. In the Efron-B method, there are n data observations and the size of the bootstrap sample is m. Let N denote the number of bootstrap samples. This bootstrap method is just sampling with replacement: For each iteration of the bootstrap, m values are sampled with replacement from n original values with equal probability to create one Efron-B sample $y = (y_1, y_2, \ldots, y_m)$. In total, N Efron-B samples are created and these bootstrap samples are used for the chosen inference.

2.3.2 Banks bootstrap

Banks [18] used linear interpolation histospline smoothing between two consecutive ordered observations when he introduced the smoothed versions of two bootstrap methods: Efron's and Bayesian (Rubin's) bootstrap. "Histospline is a smooth density estimate based only on the information in a histogram" [31]. In his paper, Banks compared those with the Bayesian bootstrap and Efron bootstrap. This thesis is interested in the smoothed Efron's bootstrap introduced by Banks, hereafter called Banks bootstrap (Banks-B), after Banks who invented it. Banks-B shows promising initial findings about its performance in quantifying the uncertainty in sample-based estimates of population characteristics for small samples [18], as discussed in Section 2.1. In the existing literature, there is work on smoothed Rubin's bootstrap [1, 129, 148], which was the main focus of Banks' paper [18]. Banks-B has not received as much attention as a more known smoothened bootstrap method, Kernel-B, which will be introduced in Appendix A.5.1. Coolen and BinHimd [53] paid further attention to Banks' version of smoothed Efron's bootstrap.

In the Banks-B method, there are n data observations and a bootstrap sample of size m is generated. The mass $1/(n+1)$ is spread uniformly between two consecutive ordered observations, $X_{(i)}$ and $X_{(i+1)}$, for $i = 1, \ldots, n+1$.

Banks and BinHimd set the left and right bounds of support for Banks-B, $x_{(0)}$ and $x_{(n+1)}$, as the minimum and maximal values of the support of the finite distribution. Banks [18] applied NPI-B to data that follow a Beta distribution, which is defined on [0, 1]. BinHimd [31], who did further investigation into Banks-B, also assumed finite support for Banks-B and in her comparison study of the bootstrap methods, she used a

Uniform distribution on a finite interval. This thesis adopts the same approach to the selection of the left and the right bounds of support for Banks-B as for NPI-B, which will be introduced in Section 2.3.3, thus, Banks-B can be applied to data with underlying distributions defined on both the infinite and finite intervals.
The Banks-B method is as follows [18]:

1. Create $n+1$ intervals from ordered n observations;
2. Sample an interval with equal probability;
3. From that interval, sample a value uniformly;
4. In total sample m values, following Steps 2 and 3, to form a Banks-B sample;
5. Create in total N Banks-B samples.

2.3.3 NPI bootstrap

NPI-B is based on $A_{(\cdot)}$, which was introduced in Section 1.4, and it is consistent with the concept of all orderings of future observations being equally likely [53]. NPI-B differs from Efron-B [78] and Banks-B [18], mainly as NPI-B was developed for prediction, while the Efron-B and Banks-B methods are aimed at quantifying the uncertainty in the estimation of population characteristics [31, 53]. In NPI-B and Banks-B, the bootstrapped observations are not restricted to already observed values. The difference between Banks-B and NPI-B is that, after sampling a value for NPI-B, this value is added to the data set before another value is sampled. This way, the number of intervals, in the partition of the part of the real-line, increases. For the first sampled value in Banks-B, the probability of it being in each interval is $\frac{1}{n+1}$, however, when another new value is added, this probability changes. For example, Banks-B is less likely to have the second sampled value in the same interval as the first sampled value and more likely to have it in a different interval. This is not the case for NPI-B. Thus, NPI-B is exactly calibrated [134, p.541]. *Exactly calibrated* means that when we simulate a model on a computer, we achieve the same proportion of events in the long run. In NPI-B, an event represents a particular combination of new points. In the frequentist theory, exact calibration is a strong consistency property. It always leads to results that are consistent with inferences based on empirical probabilities.

In the NPI-B method, there are n data observations and interest is in m future observations. Let N denote the number of bootstrap samples. In the following algorithm, sampling from the first interval, $(x_{(0)}, x_{(1)})$, and the last interval, $(x_{(n)}, x_{(n+1)})$ will be explained later. The NPI-B method is as follows [31]:

1. Take n ordered observations $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, assuming there are no ties;
2. These n observation create $n + 1$ intervals.
3. Randomly sample one of the $n + 1$ intervals, each with equal probability;
4. From that interval, sample one future value, uniformly in a finite interval;
5. Add that value to the data: increasing n to $n + 1$, and order the values;
6. Repeat Steps 2-5, now with $n + 1$ data, to get a further future value;
7. In total sample m values to form an NPI-B sample, $y = (y_1, y_2, \ldots, y_m)$, following Steps 2-6;
8. Create in total N NPI-B samples.

To apply NPI-B on the real-line, assumptions have to be made about the first interval, $(x_{(0)}, x_{(1)})$, and the last interval, $(x_{(n)}, x_{(n+1)})$, as these are often not known. Finite or infinite intervals can be selected [31]. In this thesis, five range selections are considered. For the finite bootstrap procedure, where a bounded interval is defined, $x_{(0)}$ and $x_{(n+1)}$ needs to be chosen. To do so, the left (L) and right (R) bounds of the support are selected, and $x_{(0)}$ and $x_{(n+1)}$ are set to $x_{(0)} = L$ and $x_{(n+1)} = R$, respectively. In finite bootstrap, a value is sampled uniformly from the first or the last interval. In this thesis, three approaches are considered for the finite bootstrap and then infinite, bootstrap is explored. Approach I and II are special cases of Approach III. This thesis uses the phrase half-infinite (Approach V) for infinite bootstrap employed for datasets defined on $[0, \infty)$.

I. $L = x_{(1)} - \max_i(x_{(i)} - x_{(i-1)})$ and $R = x_{(n)} + \max_i(x_{(i)} - x_{(i-1)})$, where $i = 2, 3, \ldots, n$;

II. $L = x_{(1)} - c * IQR$, $R = x_{(n)} + c * IQR$ where $c * IQR$ is the interquartile range (IQR) of the original dataset multiplied by a constant $c > 0$;

III. $L = x_{(1)} - v$, $R = x_{(n)} + v$ where $v > 0$ is a constant;

IV. Infinite bootstrap: There are $n+1$ intervals created by n observations. For intervals between $x_{(1)}$ and $x_{(n)}$, the same procedure is used as for finite intervals. To sample from $(-\infty,\ x_{(1)})$ and $(x_{(n)},\ \infty)$, Normal distribution tails fitted to the intervals are assumed, with estimated mean $\mu = \frac{x_{(1)}+x_{(n)}}{2}$ and estimated standard deviation $\sigma = \frac{x_{(n)}-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ denotes the cumulative of the standard Normal distribution [31]. σ is estimated using the properties of the Normal cumulative function: $P(Y > x_{(n)}) = 1 - \Phi\big(\frac{x_{(n)}-\mu}{\sigma}\big) = \frac{1}{n+1}$ [31].

V. Half-infinite bootstrap: $L = 0$, a value from the last interval, $(x_{(n)}, \infty)$, is sampled by assuming tails of an Exponential distribution. To estimate the parameter of Exponential distribution, λ, the cumulative function $P(Y < y) = 1 - e^{(-\lambda y)}$ is used. Given that $P(Y > x_{(n)}) = \frac{n}{n+1}$, the parameter $\lambda = \frac{\ln(n+1)}{x_{(n)}}$ is estimated [31].

This thesis explores all five ways to define the range in this chapter and in Chapter 4.

2.3.4 Hutson bootstrap

Hutson [111] introduced a new quantile function estimation method for generating bootstrap samples: the semi-parametric composite quantile function estimator, which combines a parametric model with a standard linear interpolation quantile function estimator. Quantile function specifies the value of a random variable, given the chosen probability, in such a way that the probability of the variable is less than or equal to that value. Although this method was not formulated by Hutson [111] as a bootstrap method, but rather as a function for generating bootstrap samples, for simplicity, this thesis calls it Hutson bootstrap (Hutson-B). Hutson [111] showed that Hutson-B improves the coverage probabilities of the standard bootstrap percentile confidence intervals. Hutson-B allows ties [111]. Two applications of Hutson-B in pharmaceutical research were introduced in Section 2.2.

Let $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ be ordered observation from an i.i.d sample of size n from a continuous distribution F defined on the real line. Then Equation (2.2) presents the semi-parametric composite quantile function estimator, as defined by Hutson [111].

In what follows, $\lfloor \cdot \rfloor$ is the floor function, $n' = n + 1$, ϵ is defined in Equation (2.1) and $0 < u < 1$.

$$\epsilon = n'u - \lfloor n'u \rfloor \tag{2.1}$$

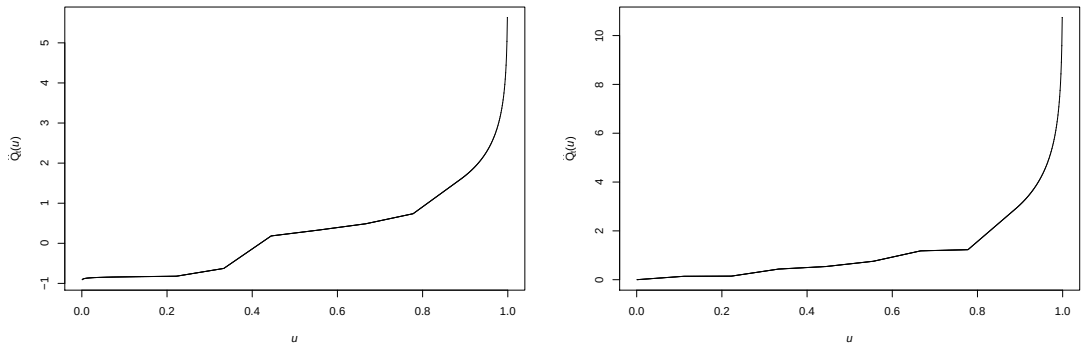
$$\hat{Q}_T(u) = \begin{cases} x_{(1)} + (x_{(2)} - x_{(1)}) \log((n+1)u), & \text{if } 0 < u \leq \frac{1}{1+n} \\ \hat{Q}_L(u) = (1-\epsilon) x_{(\lfloor (n+1)u \rfloor)} + \epsilon x_{(\lfloor (n+1)u \rfloor + 1)}, & \text{if } \frac{1}{1+n} < u < \frac{n}{1+n} \\ x_{(n)} - (x_{(n)} - x_{(n-1)})) \log((n+1)(1-u)), & \text{if } \frac{n}{1+n} \leq u < 1 \end{cases} \tag{2.2}$$

$\hat{Q}_L(u)$ in Equation (2.2) represents the standard linear interpolation quantile function estimator. Hutson-B, as defined in Equation (2.2), is used for distributions defined on $(-\infty, \infty)$, such as Normal and Mixed-Normal distribution. For distributions defined on $(0, \infty)$, such as Lognormal and Exponential distributions, Equation (2.3) is used instead. $\hat{Q}_T(u)$ depends on the data sample. In Figure 2.1, an illustration of plots of $\hat{Q}_T(u)$ is provided for two simple data samples of sample size $n = 8$, $a \sim$ N(0,1) and $b \sim$ Exp(1), where a=(-0.836, -0.820, -0.626, 0.184, 0.330, 0.487, 0.738, 1.595) and b=(0.140, 0.146, 0.436, 0.540, 0.755, 1.182, 1.230, 2.895).

$$\hat{Q}_T(u) = \begin{cases} \epsilon x_{(1)}, & \text{if } 0 < u \leq \frac{1}{1+n} \\ \hat{Q}_L(u) = (1-\epsilon) x_{\lfloor (n+1)u \rfloor} + \epsilon x_{\lfloor (n+1)u \rfloor + 1}, & \text{if } \frac{1}{1+n} < u < \frac{n}{1+n} \\ x_{(n)} - (x_{(n)} - x_{(n-1)}) \log((n+1)(1-u)), & \text{if } \frac{n}{1+n} \leq u < 1 \end{cases} \tag{2.3}$$

The following algorithm creates m Hutson bootstrap values from the n original values:

1. Take n ordered observations $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ from an i.i.d. sample of size n;
2. Generate a random sample of size m from the standard uniform distribution;
3. Apply the semi-parametric composite quantile function estimator $\hat{Q}_T(u)$ to the m values generated in Step 2, forming an Hutson-B sample;
4. Create in total N Hutson-B samples.



(a) Sample $a \sim$ N(0,1), $n_a = 8$ (b) Sample $b \sim$ Exp(1), $n_b = 8$

Figure 2.1: Plots of $\hat{Q}_T(u)$ for two simple data samples

Hutson [111] compared Hutson-B [111] to other two quantile function estimators: Hutson and Ernst's [110] sample quantile function estimator and Harrel-Davis kernel quantile function estimator (see Sheather and Marron [186]). The details regarding these quantile functions are outside the scope of this thesis. For $n = 10, 25$ he compared coverage probability at $\alpha = 0.05$ for a variety of distributions (Normal, Logistic, Laplace, Cauchy, Exponential, half-Normal and Rayleigh) and for the estimation of various statistics (mean, median, standard deviation, skewness, excess kurtosis, Q3 and upper decile). He concluded that the performance of Hutson-B is superior to the other two quantile function estimators, the biggest difference was apparent at $n = 10$. Hutson [111] acknowledged a limitation of Hutson-B: that it is unable to fully capture the tail behaviour of heavy-tailed distributions, such as of Cauchy distribution.

Hutson-B has been applied in hydrology, particularly in extrapolation of hydrological extremes. Jagtap et al. [123] further developed on Hutson-B and compared it with non-parametric and parametric bootstrap, focusing on coverage probability for small samples, when estimating high hydro-meteorological quantiles and extreme events. Sample sizes explored in the simulation study were $n = 10, 25, 50, 100$. An example of precipitation dataset with $n = 25, 27, 37, 37, 84, 117$ was given.

Further work has been done by Hutson [112], who developed a sigmoidal quantile function estimator and a hybrid quantile function estimator. The latter combines the properties of the kernel quantile function with the sigmoidal quantile function estima-

tor. Hutson [112] argued that the generalised sigmoidal quantile function "can estimate quantiles beyond the range of the data, which is important for certain applications given smaller sample sizes" [112]. The study of those variations of Hutson-B is outside the scope of this thesis and a topic for future research.

2.3.5 Coverage and bootstrap confidence intervals

This chapter presents a study which assesses both the estimation and prediction performances by focusing on coverage of confidence intervals. The term coverage refers to the proportion of the times in the long run that a confidence interval contains the true value of interest. Ideally, coverage should equal the confidence level. A $(1-2\alpha)100\%$ confidence interval (CI) is used when estimating an unknown parameter from a sample. The $(1-2\alpha)100\%$ confidence interval, $(\hat{\theta}^{(\alpha)}, \hat{\theta}^{(1-\alpha)})$, can be written as Equation (2.4), where $\hat{\theta}^{(\alpha)}$ and $\hat{\theta}^{(1-\alpha)}$ are both functions of the data X, the confidence level is $(1-2\alpha) \in [0,1]$ and it does not depend on θ. If 90% confidence intervals are formed for chosen population parameter θ for 100 samples, 90 of these confidence intervals are expected to include the true value of θ.

$$P(\hat{\theta}^{(\alpha)} \leq \theta \leq \hat{\theta}^{(1-\alpha)}) = 1-2\alpha \tag{2.4}$$

A confidence interval can be calculated for a chosen population characteristic of interest, such as mean, median, variance, Q1, Q3, or IQR. This section will introduce percentile confidence intervals and BC_a confidence intervals, both of which will be employed in Section 2.4. Percentile confidence intervals have also been called quantile confidence intervals in the literature, however, this thesis will use the term percentile confidence intervals.

To calculate a confidence interval for a chosen population characteristic of interest, θ, the following inputs are important: B independent bootstrap samples $y^{*1}, y^{*2}, \ldots, y^{*B}$, each of size m, and the bootstrap replication of $\hat{\theta}$ corresponding to each bootstrap sample $\hat{\theta}^*(b) = s(y^{*b})$, for $b \in \{1, 2 \ldots, B\}$. Here s represents the formula calculating a particular sample statistic. The $\hat{\theta}^*$s are ordered in an ascending order. For BC_a confidence intervals, the knowledge is required about the estimate of θ based on the observed data of the original sample, $\hat{\theta} = s(x)$, and the original sample, $x = (x_1, x_2, \ldots, x_n)$. For example, if θ

is the true population mean, $\hat{\theta} = s(x) = \frac{1}{n}\sum_i^n x_i$ is the sample mean, i.e. estimate of the true population mean based on the original sample.

The $100(1-2\alpha)$ percentile confidence interval $(\hat{\theta}^{*(\alpha)}, \hat{\theta}^{*(1-\alpha)})$ is computed by taking the $(B\alpha)^{th}$ and $(B(1-\alpha))^{th}$ value of the ordered $\hat{\theta}^*$s. $\hat{\theta}^{*(\alpha)}$ stands for the 100αth percentile of B $\hat{\theta}^*$s. For example, 90% percentile confidence interval, i.e. at $\alpha = 0.05$, is the interval $(\hat{\theta}^{*(50)}, \hat{\theta}^{*(950)})$.

As stated in Section 2.1, Efron has improved the percentile confidence intervals, to account for and correct the bias and the skewness of the bootstrap parameter estimate. The two improved versions of the Efron-B confidence intervals are the bias-corrected and accelerated bootstrap (BC_a) and the approximate bootstrap confidence (ABC) intervals [69]. ABC analytically approximates the BC_a interval endpoints [69] and it is less computationally demanding compared to BC_a. Given ABC and BC_a similarity, the performance in estimation for small samples ($n = 4, 6, 8, 10$) is studied only for BC_a intervals. An advantage of BC_a confidence intervals over percentile confidence intervals named in the literature is their high order of accuracy [31].

In order to calculate BC_a confidence intervals, the bias-correction $\hat{z}_0$ and the acceleration $\hat{a}$ need to be computed first. The bias-correction $\hat{z}_0$, calculated via Equation (2.5), is based on the $\hat{\theta}^*(b)$s and the original sample estimate, $\hat{\theta}$. Thus, $\hat{z}_0$ is influenced by the choice of the bootstrap method.

$$\hat{z}_0 = \Phi^{-1}\left(\frac{\#\{\hat{\theta}^*(b) < \hat{\theta}\}}{B}\right) \tag{2.5}$$

The acceleration $\hat{a}$ adjusts the skewness of the bootstrap distribution. The acceleration $\hat{a}$ is based on the original sample, not the bootstrap samples, and it can be computed in multiple ways. This thesis employs the method described by Efron and Tibshirani [78], which utilises jackknife values of a statistic $\hat{\theta} = s(x)$. Let $x_{(-i)}$ be the original sample with the ith point x_i deleted. Now $\hat{\theta}_{(-i)} = s(x_{(-i)})$ and $\hat{\theta}_{(\cdot)}$ is defined in Equation (2.6).

$$\hat{\theta}_{(\cdot)} = \sum_{i=1}^{n} \hat{\theta}_{(-i)}/n \tag{2.6}$$

Then $\hat{a}$ is defined in Equation (2.7).

$$\hat{a} = \frac{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)})^3}{6\{\sum_{i=1}^n (\hat{\theta}_{(\cdot)} - \hat{\theta}_{(-i)})^2\}^{\frac{3}{2}}} \tag{2.7}$$

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{\alpha}}{1 - \hat{a}(\hat{z}_0 + z^{\alpha})}\right) \tag{2.8}$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{1-\alpha}}{1 - \hat{a}(\hat{z}_0 + z^{1-\alpha})}\right) \tag{2.9}$$

The $(1-2\alpha)$ BC_a interval is $(\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)})$; α_1 and α_2 can be calculated via Equations (2.8) and (2.9), respectively. R functions were written for implementation.

When calculating BC_a intervals for Efron-B for small samples for the estimation of quantiles and IQR for Normally distributed data (see Section 2.4.2) and for the estimation of most statistics for Lognormally distributed data (see Section 2.4.3), a problem arose for Efron-B. The problem was that in some cases, the output of $\hat{\theta}^{*(\alpha_1)}$ was NA. This happened when $\alpha_1 < 0.001$. At $\alpha_1 = 0.001$, $\hat{\theta}^{*(\alpha_1)}$ is the smallest bootstrap sample statistic, given $N = 1000$, and below $\alpha_1 = 0.001$, $\hat{\theta}^{*(\alpha_1)}$ is undefined. This problem most likely occurred because Efron bootstrap samples only contain values from the original observations and in cases where the sample size of the original sample is small, there are very few values to sample from. The problem was fixed in the R code by setting $\alpha_1 = 0.001$ if $\alpha_1 < 0.001$.

Bootstrap-t confidence intervals have also been used in research [20]. In some sources, they are also called the Student's t method [127] and the percentile method with a studentised pivot [18]. Bootstrap-t confidence intervals assume that $\frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{\text{se}}(b)}$ is approximately t-distributed [127]. $(1 - 2\alpha)\%$ bootstrap-t confidence intervals at a certain α level is $\left(\hat{\theta} - t_{n-1}^{(1-\alpha)}\widehat{\text{se}}, \hat{\theta} - t_{n-1}^{(\alpha)}\widehat{\text{se}}\right)$ [78]. Bootstrap-t confidence intervals are suitable for the estimation of confidence intervals for location statistics, such as the sample mean [69]. However, bootstrap-t confidence intervals can be unpredictable for small samples and in nonparametric situations [69]. Despite this, Banks [18] explored bootstrap-t confidence intervals in his simulation study of the performance of Banks-B for small samples. Further discussion about the suitability of bootstrap-t confidence intervals is outside the scope of this thesis and bootstrap-t confidence intervals are not included in this simulation study.

To summarise, there is no consensus on the recommendation of a particular confidence interval for small samples and the scope of this thesis is limited to percentile and BC_a confidence intervals.

2.4 Bootstrap coverage performance in estimation

This study assesses how well the four bootstrap methods (Efron-B, Banks-B, Hutson-B and NPI-B) perform at quantifying the uncertainty in the estimation of population characteristics (mean, variance, median, Q1, Q3 and IQR), with a focus on small sample sizes ($n = 4, 6, 8, 10$). NPI-B [31, 53] is aimed at prediction, not estimation. NPI-B is included in this study to explore whether in some cases it can provide good confidence intervals in the estimation inference.

There are two bootstrap method comparison studies that inspired this further investigation into the bootstrap coverage performance in quantifying the uncertainty in the estimation of population characteristics for small samples. The first one is Banks' comparison of Banks-B, smoothened Bayesian bootstrap, Efron-B and Bayesian bootstrap [18]. Banks [18] looked at the estimation of the mean, median and variance, considering three sample sizes: $n = 5, 10, 20$. Note that for the estimation of median he used $n = 6$ instead of $n = 5$ for computational simplicity. The data in his study come from Beta distributions with different parameters. Banks [18] used a goodness-of-fit test to compare confidence regions, which will be also carried out in this section.

The second study is BinHimd's comparison of the bootstrap performance in estimation [31]. BinHimd [31] compared three bootstrap methods, Efron-B, Banks-B and NPI-B on finite intervals, using Uniform and Beta distributions. For each bootstrap method, BinHimd [31] generated $B = 1000$ bootstrap samples, she calculated a chosen statistic for each bootstrap sample, $\hat{\theta}^*(b)$, $b \in \{1, \ldots, B\}$, and then she calculated the variance, bias, absolute error, and mean square error of $\hat{\theta}^*(b)$s for the three bootstrap methods, focusing on the estimation of mean, variance and Q3, for sample sizes $n = 20, 50, 100, 200, 500, 1000$. Bias, mean square error and absolute error are commonly used measures of statistical accuracy of estimators [78]. On the real-line, BinHimd compared only Efron-B and NPI-B, using Uniform, Normal and Gamma distributions, and sample sizes $n = 20, 50, 100, 200, 500$; apart from the same analysis, she also considered coverage of 90% and 98% confidence intervals for the two bootstrap methods. BinHimd used BC_a confidence intervals instead of percentile confidence intervals.

The limitations of Banks' and BinHimd's work is that in their exploration of Banks-B, they focused on data generated from distributions with finite support. Infinite support

has been explored only for NPI-B, not for Banks-B. This study further explores Banks-B and NPI-B in cases where there is infinite support and for distributions that BinHimd or Banks did not consider, i.e. Mixed-Normal, Exponential and Lognormal. BinHimd did not consider sample sizes below $n = 20$ in her comparison study of the bootstrap methods performance in the estimation of population characteristics. Hutson-B has never been compared to Banks-B and NPI-B. This study extends the exploration, focusing on smaller samples, $n = 4, 6, 8, 10$, and by including Hutson-B in the comparison study.

The methodology has been inspired by Banks [18] and Al Luhayb [4]. Banks [18] introduced the algorithm for the bootstrap method comparison based on chi-square goodness of fit test and he considered 20 and 100 confidence regions.. Later Al Luhayb [4] adopted the algorithm to compare the generalised Banks' smoothed bootstrap method and Efron-B for right-censored data. Al Luhayb considered 10 confidence regions. Al Luhayb [4] considered sample sizes $n = 6, 10, 20, 40, 100$, he focused on the estimation of Q1, Q2 and Q3 and he based his analysis on the χ^2-value. He concluded that the generalised Banks' bootstrap performed well in the estimation of the first, second and third quartiles. The main difference between the algorithm employed in this thesis and the simulations from which the algorithm was adopted is that there are several runs carried out in the algorithm to improve the robustness of the conclusions, whereas Banks [18] and Al Luhayb [4] presented results from only one run. The outputs of the simulation study introduced in this thesis are presented in boxplots rather than in tables to allow for a visual comparison of the bootstrap methods. Banks [18] also compared the algorithm outputs for the studentised confidence intervals. However, the consideration of these confidence intervals is outside the scope of this thesis.

In Section 2.4.1, the algorithm for the evaluation of the bootstrap performance in estimation is presented, and two metrics for the evaluation of the performance are introduced. The algorithm is applied to data from four different distributions: Normal, Exponential, Lognormal and Mixed-Normal. The simulation outputs are presented in three parts: Section 2.4.2 discusses observations for data generated from Normal distribution; Section 2.4.3 analyses conclusions for data generated from Lognormal and Exponential distributions; and Section 2.4.4 examines findings for data generated from Mixed-Normal distribution. Finally, Section 2.4.5 summarises the findings of the study of the bootstrap performance

in estimation.

2.4.1 Methodology

The simulation for evaluation of the performance in estimation for the bootstrap methods is described in Algorithm 1. In this investigation, $N = 1000$, $B = 1000$, $m = n$ and $M = 20$. M present the number of runs of the algorithm. The choice of using $B = 1000$ is in alignment with other work on bootstrap confidence intervals (see Efron and Tibshirani [78]). The main focus is on sample sizes $n = 4, 6, 8, 10$. Unless stated otherwise, $(1-2\alpha)\%$ percentile confidence intervals are calculated in Algorithm 1.

To assess the performance of bootstrap methods, this investigation focuses on two metrics generated through simulations: the coverage at 90% confidence interval (CI) and the χ^2-value arising from a chi-square goodness of fit test on the confidence regions. The chi-square goodness of fit test tests the hypothesis that all confidence regions have equal coverage probabilities for the given statistic of interest [18]. Both metrics are important for the bootstrap method assessment of its performance in the estimation of population characteristics. Thus, the conclusions are based on both metrics.

The first metric used to assess the performance in the estimation of a particular population characteristic is the coverage at 90% CI. There are three types of coverage: under-coverage (below 90%), good coverage (around 90%) or over-coverage (above 90%). The best coverage at 90% CI is at (or close to) 90% as this shows that the bootstrap method provides the expected estimate at this level. Over-coverage means that the method provides more precise estimates than it should whereas under-coverage means that the method provides less precise estimates than it should. Under-coverage is worse than over-coverage because over-coverage at least guarantees the performance in the estimation of population characteristics and it is risk-free whereas under-coverage is not informative about the accuracy of the estimation. While over-coverage sounds ideal, it is not because the confidence intervals are too wide. Brief investigation into coverage at 95% CI for the estimation of mean and variance showed same pattern for both coverage at 90% and 95% CI, but there is more under-coverage at 95% CI. Thus, reporting just 90% is a reasonable action.

Algorithm 1 Bootstrap performance in the estimation of population parameter θ

1: Generate N data sets $\{x_k^i\}_{k=1}^n$, $i \in 1, \ldots, N$, of sample size n from the chosen population;

2: For each data set, apply the bootstrap method to generate B bootstrap samples of size m, $\{y_k^{ib*}\}_{k=1}^m$, and calculate the statistic of interest for each of those bootstrap samples, $\hat{\theta}_i^*(b)$, $b \in \{1, 2 \ldots, B\}$;

3: For each data set, compute $(1-2\alpha)\%$ confidence intervals $(\hat{\theta}_i^{*(\alpha)}, \hat{\theta}_i^{*(1-\alpha)})$ for θ at different confidence levels ($\alpha = 0.05, 0.1, \ldots, 0.45$);

4: For each data set and for each confidence level, record whether $\theta \in (\hat{\theta}_i^{*(\alpha)}, \hat{\theta}_i^{*(1-\alpha)})$;

5: For each confidence level, calculate the proportion of confidence intervals for which $\theta \in (\hat{\theta}_i^{*(\alpha)}, \hat{\theta}_i^{*(1-\alpha)})$: $\rho_\alpha = \frac{1}{N}\sum_{i=1}^N \mathbb{1}\{\theta \in (\hat{\theta}_i^{*(\alpha)}, \hat{\theta}_i^{*(1-\alpha)})\}$.

6: Calculate the observed statistics, $O_{(j)}$, for 10 confidence region, $j = 1, 2, \ldots, 10$, across all N. $O_{(j)} = \sum_{i=1}^N \mathbb{1}\{\theta \in CR_{(j)}^i\}$. $O_{(j)}$ can be calculated from Step 5, as follows: $O_{(1)} = N\rho_{0.45}$, $O_{(2)} = N(\rho_{0.4} - \rho_{0.45})$, $O_{(3)} = N(\rho_{0.35} - \rho_{0.4}), \ldots, O_{(9)} = N(\rho_{0.05} - \rho_{0.1})$, $O_{(10)} = N(1 - \rho_{0.05})$.

7: Carry out the chi-square goodness of fit test on $O_{(j)}$ and record the χ^2-value and the coverage (in %) at 90% confidence interval;

8: Repeat Steps 2-7 M times in total. For each run, report the coverage at 90% confidence interval (in %), i.e. $100\rho_{0.05}$, and the χ^2-value.

The second metric of interest is the χ^2-value calculated via the chi-square test. The added value of using χ^2-value alongside coverage at 90% CI is that it assesses "the discrepancy in coverage probability" [18]. The chi-square test considers 10 confidence regions with their observed values $O_{(j)}$ for $j = 1, 2, \ldots, 10$. Confidence region is defined as follows [4]:

$$CR_{(j)} = (\hat{\theta}^{*(\frac{\alpha_{j+1}}{2})}, \hat{\theta}^{*(\frac{\alpha_j}{2})}) \cup (\hat{\theta}^{*(1-\frac{\alpha_j}{2})}, \hat{\theta}^{*(1-\frac{\alpha_{j+1}}{2})}), \alpha_1 = 1, \alpha_{j+1} = \alpha_j - 0.1 \tag{2.10}$$

In Algorithm 1, the nominal coverage probability of each confidence region is set at 0.10. Thus, $CR_{(1)} = (\hat{\theta}^{*0.45}, \hat{\theta}^{*0.5}) \cup (\hat{\theta}^{*0.5}, \hat{\theta}^{*0.55}) = (\hat{\theta}^{*0.45}, \hat{\theta}^{*0.55})$, $CR_{(2)} = (\hat{\theta}^{*0.4}, \hat{\theta}^{*0.45}) \cup (\hat{\theta}^{*0.55}, \hat{\theta}^{*0.6})$, $CR_{(3)} = (\hat{\theta}^{*0.35}, \hat{\theta}^{*0.4}) \cup (\hat{\theta}^{*0.6}, \hat{\theta}^{*0.65})$, $\ldots$, $CR_{(10)} = (\hat{\theta}^{*0}, \hat{\theta}^{*0.05}) \cup (\hat{\theta}^{*0.95}, \hat{\theta}^{*1})$. The actual coverage at each confidence region can be calculated from $(1-2\alpha)\%$ confidence intervals at different α levels, $\alpha = 0.05, 0.1, \ldots, 0.45$. This is done in the following way:

$\theta \in CR_{(1)}$ iff (if and only if) $\{\theta \in 10\%$ CI$\}$, $\theta \in CR_{(2)}$ iff $\{\theta \in 20\%$ CI$\} \cap \{\theta \notin 10\%$ CI$\}$, $\theta \in CR_{(3)}$ iff $\{\theta \in 30\%$ CI$\} \cap \{\theta \notin 20\%$ CI$\}$, $\ldots$, $\theta \in CR_{(10)}$ iff $\{\theta \notin 90\%$ CI$\}$.

The expected value for each confidence region is $E_j = N/10$ and $\chi^2 = \sum_{j=1}^{10} \frac{(O_j - E_j)^2}{E_j}$. The χ^2-statistic compares the size of any discrepancies between the expected values and the actual values, given the number of data sets, N, and the number of confidence regions considered, i.e. 10 in Algorithm 1. The statistic χ^2 has a Chi-squared distribution with 9 degrees of freedom and H_0 is rejected when $\chi^2 > 16.919$ (at $\alpha = 0.05$). The lower the χ^2-value is, the better the bootstrap method performs. When the χ^2-values become large, then it is clear that the bootstrap method performs poorly. It is important to check whether a high χ^2-value is mainly caused by under-coverage or over-coverage. Thus, the χ^2-value and coverage at 90% CI are considered together to arrive at an assessment of the performance of a bootstrap method in the estimation of a particular population characteristic.

A small simulation was run to study what would be the effect on the χ^2-value of increasing $N = 1,000$ to $N = 10,000$, and of increasing the number of confidence regions from 10 to 20. In both cases, the prevailing patterns and comparisons between different bootstrap methods remain similar. When N is set to $10,000$ instead of $1,000$, the squared differences between O_j and E_j become much larger, by a roughly factor of 10^2 each, while the denominator only becomes larger by a factor of 10. Thus, the cell observations are about 10 times larger, with random variation, meaning the χ^2-value is roughly 10 times larger for $N = 10,000$ than for $N = 1,000$. Clearly, increasing the number of data emphasises the discrepancy between the null hypothesis of equal numbers per cell and the actual case. Using 20 instead of 10 confidence regions in Algorithm 1 did not affect the conclusions made regarding bootstrap method comparison. Efron-B performs notably worse when 20 instead of 10 confidence regions are used in Algorithm 1, i.e. it has higher χ^2-value, compared to the other bootstrap, however the conclusions made regarding bootstrap method comparison remain the same.

2.4.2 Normally distributed data

This section presents findings on the bootstrap method performance in the estimation of population characteristic for Normally distributed data. The data are simulated from Normal distribution with mean $\mu = 1$ and standard deviation $\sigma = 1$. Because of the standardising properties of Normal distributions, the performance in estimation is not affected by the choice of parameters. If different parameters were chosen and the simulation study was carried out, the same patterns would be seen. For the Normal distribution, the true population characteristics calculations are: Q1 for $N(\mu, \sigma^2)$ equals to $\mu - 0.67448\sigma$ and Q3 for $N(\mu, \sigma^2)$ equals to $\mu + 0.67448\sigma$. Thus, true Q1 for $N(1, 1^2)$ is 0.32552 and Q3 for $N(1, 1^2)$ is 1.67448.

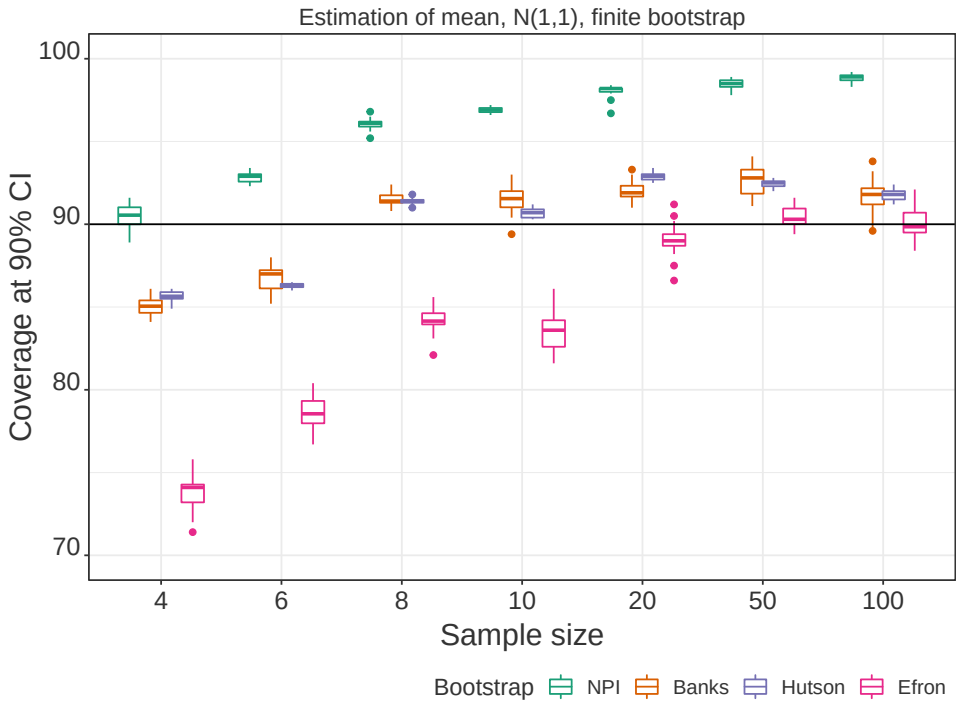
For Banks-B and NPI-B, finite range (Approach I, Section 2.3.3) is assumed and findings on the estimation of population characteristics are presented. The influence of the tails assumption for these two bootstrap methods will be considered later.

Brief consideration of larger samples

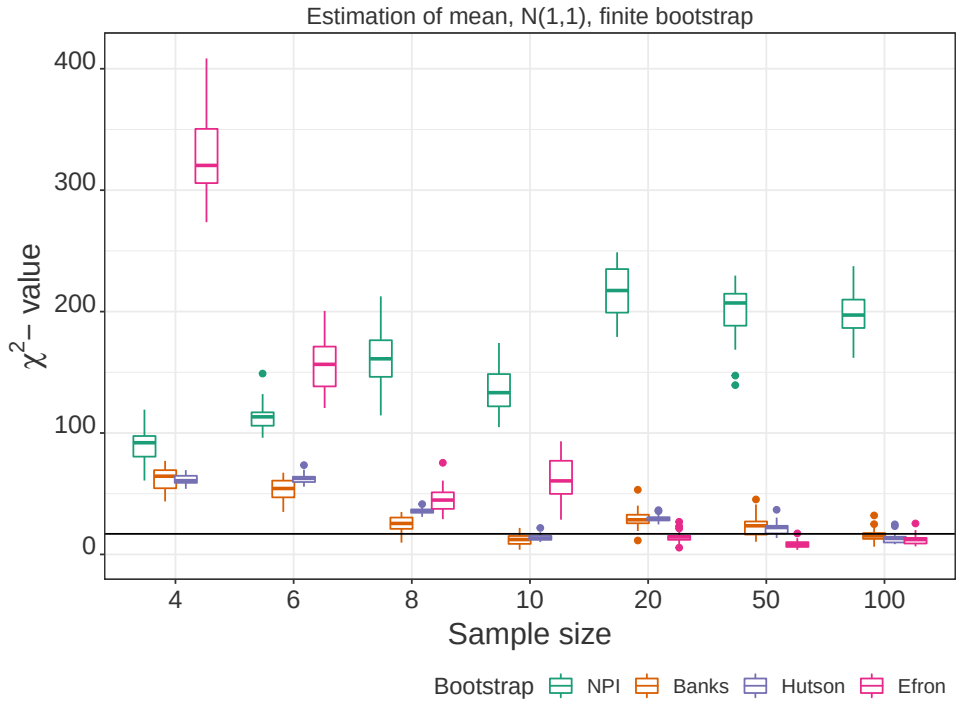
This study of the bootstrap performance in estimation is primarily concerned with small sample sizes (up to $n = 10$). Nevertheless, for the estimation of mean (Figure 2.2) and variance (Figure 2.3), the simulations have been carried out for a larger spectrum of sample sizes, $n = 4, 6, 8, 10, 20, 50, 100$. The inclusion of larger sample sizes has shown a pattern: the Efron-B performance improves as n increases, i.e. χ^2-value decreases and the coverage at 90% CI improves as n increases. Efron-B has a very good performance in the estimation of mean from $n = 20$: χ^2-value is low and the coverage at 90% CI is close to the ideal coverage. In the estimation of variance, Efron-B is the best performing bootstrap only for $n = 100$. These findings are consistent with the asymptotic theory [63] that is a theoretical justification for Efron-B [63], as discussed in Section 2.1. As the sample size n increases, the sample becomes more representative of the real underlying population distribution. The rest of this chapter will be devoted to smaller sample sizes.

Estimation of mean

For the estimation of mean for sample sizes $n = 4, 6, 8, 10$ for Normally distributed data (part of Figure 2.2), Banks-B and Hutson-B are the best performing bootstrap methods

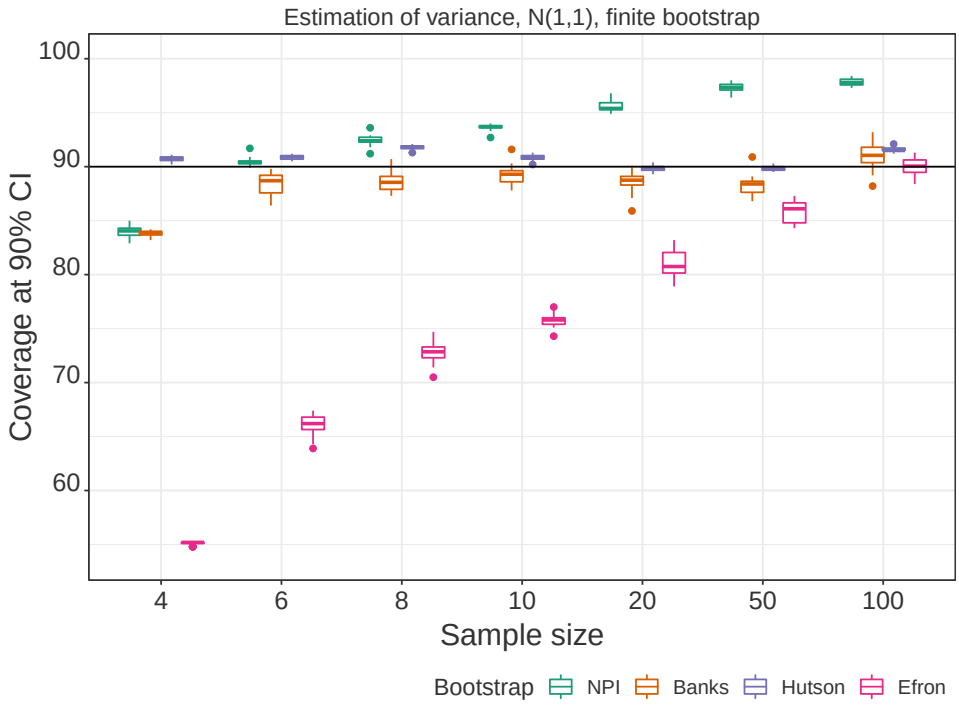


(a)

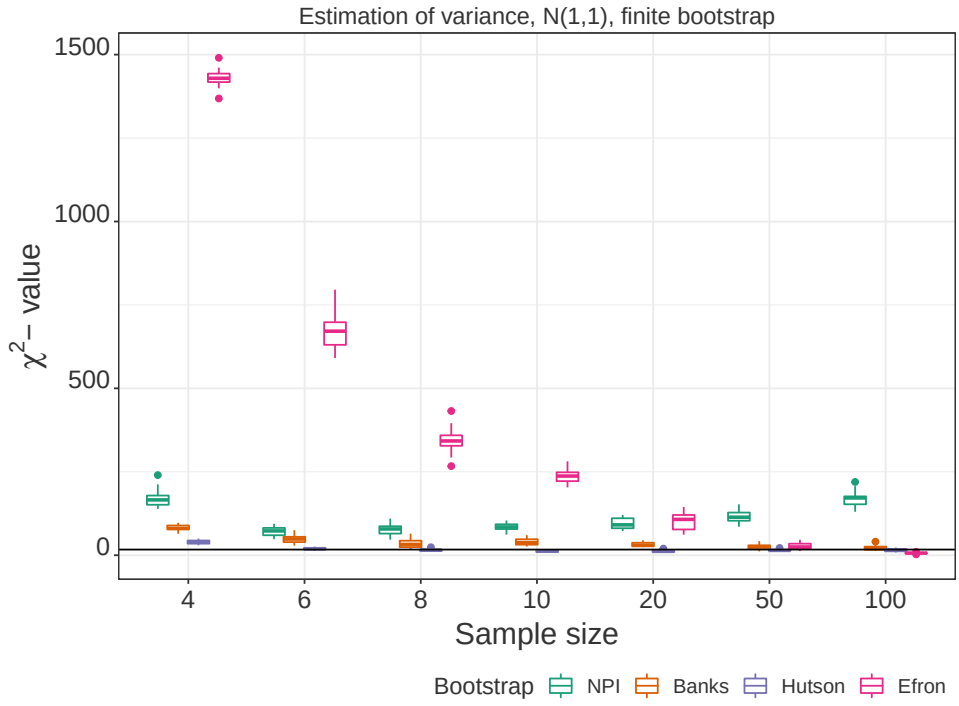


(b)

Figure 2.2: Coverage at 90% CI and χ^2-values, estimation of mean, N(1,1), $n = 4, 6, 8, 10, 20, 50, 100$, finite (Approach I) Banks-B and NPI-B, 20 simulations



(a)



(b)

Figure 2.3: Coverage at 90% CI and χ^2-values, estimation of variance, N(1,1), $n = 4, 6, 8, 10, 20, 50, 100$, finite (Approach I) NPI and Banks-B, 20 simulations

from the perspective of both metrics of assessment. Banks-B is marginally better for $n = 6, 8, 10$ because it has the lowest χ^2-values for these sample sizes. For both Hutson-B and Banks-B, there is over-coverage at 90% CI for $n = 8, 10$ rather than under-coverage. The performance in the estimation of mean for NPI-B is the best at $n = 4$ where there is ideal coverage at 90% CI and the χ^2-values for NPI-B are the lowest, however, even at this case, the χ^2-values for NPI-B are larger than for Banks-B and Hutson-B. There is very large variation of simulation outputs for Efron-B at $n = 4$ and $n = 6$. There are some discrepancies between Figures 2.2 (a) and 2.2 (b): For $n = 4$, the coverages at 90% CI are differently ordered for NPI-B, Hutson-B and Banks-B to the inaccuracies in χ^2-values - whereas NPI-B has the best coverage at 90% CI, it has the third best χ^2-value; Hutson-B has the second best coverage, but the best χ^2-value; and Banks-B has the third best coverage, but the second best χ^2-value. For $n = 8, 10$ Banks-B has lower χ^2-value than Hutson-B, but Hutson-B has coverage that is slightly closer to ideal coverage. As both Hutson-B and Banks-B have only small over-coverage, this does not affect the conclusions. Overall, Banks-B and Hutson-B perform consistently well in the estimation of mean for Normally distributed data with small sample sizes ($n = 10$ and smaller).

Estimation of variance

Hutson-B is the best performing bootstrap in the estimation of variance for Normally distributed data for small sample sizes (part of Figure 2.3). Hutson-B has consistently small over-coverage at 90% CI and it has the lowest χ^2-value. Banks-B has the second lowest χ^2-value and the third best coverage at 90% CI (small under-coverage). NPI-B has the second best coverage at 90% CI and the third lowest χ^2-value. Given that Banks-B has only small under-coverage, this thesis concludes that it is the second best bootstrap method in the estimation of variance for Normally distributed data for small sample sizes.

Estimation of quantiles and IQR

The bootstrap performance in the estimation of sample quantiles and IQR will be discussed henceforth. There are more ways to calculate sample quantiles and IQR [114]. The default type in R, Type 7, is used. Investigation into how the choice of quantile types affects the performance in the estimation of quantiles for Normally distributed data

has been carried out and initial remarks, alongside definitions of different quantiles types, can be found in Appendix A.1. Using a different type of quantile calculation can make a difference on the value of the sample statistic for small samples and the performance of the bootstrap method is slightly affected by the choice of the quantile types. The most affected bootstrap method is Efron-B. The findings about the bootstrap method performance in the estimation of median are consistent across different ways of calculating quantiles. When estimating Q1 and Q3, the choice of quantile type can affect whether Hutson-B or Banks-B performs better. However, given that both bootstrap methods perform well in such cases, this is not a big concern.

Clearly, Efron-B is the worst performing bootstrap in the estimation of quantiles for Normally distributed data for small sample sizes, as it has under-coverage for $n = 4, 6, 8, 10$ when quantiles (Q1, median, Q3) are estimated, except for $n = 10$ in the estimation of Q3.

When estimating Q1, at $n = 4$, NPI-B has the best coverage at 90% CI (close to the ideal coverage), Banks-B has under-coverage (but still the second best coverage) and it has the lowest χ^2-value. At $n = 4$, Hutson-B has under-coverage and the second lowest χ^2-value. At $n = 6, 8$, Banks-B is the best performing bootstrap method in the estimation of Q1 for Normally distributed data from the perspective of both metrics of assessment and at $n = 10$, Banks-B and Hutson-B both perform equally well (almost ideal coverage at 90% CI and low χ^2-values).

In the estimation of median, Banks-B has small under-coverage at $n = 4$ but its coverage is close to the ideal coverage. NPI-B has over-coverage, which is preferable to under-coverage at $n = 4$, but it has higher χ^2-value compared to Banks-B and Hutson-B. From $n = 6$, both Banks-B and Hutson-B have good coverage at 90% CI and low χ^2-value, Hutson-B has slightly lower χ^2-value. This work concludes that for $n = 4$, Banks-B is the best performing bootstrap method in the estimation of median for Normally distributed small samples, and for $n = 6, 8, 10$, Hutson-B performs the best.

In the estimation of Q3, for $n = 4$, NPI-B has coverage at 90% CI that is the closest to the ideal coverage (there is still small under-coverage), but Banks-B has the lowest χ^2-value. For $n = 6$, Banks-B has the best coverage at 90% CI (small under-coverage, close to the ideal coverage) and the smallest χ^2-value. At $n = 8, 10$, Banks-B and Hutson-B

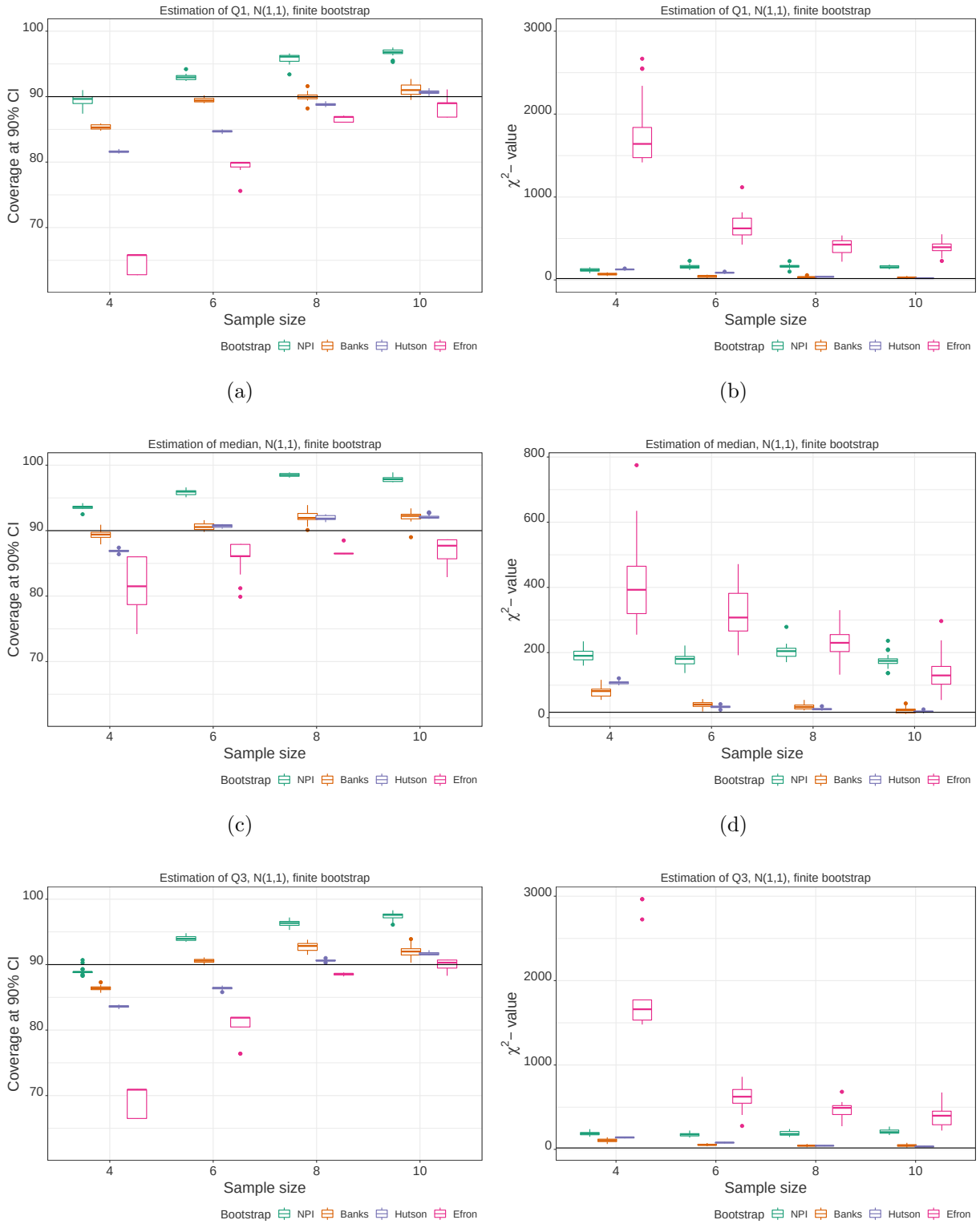
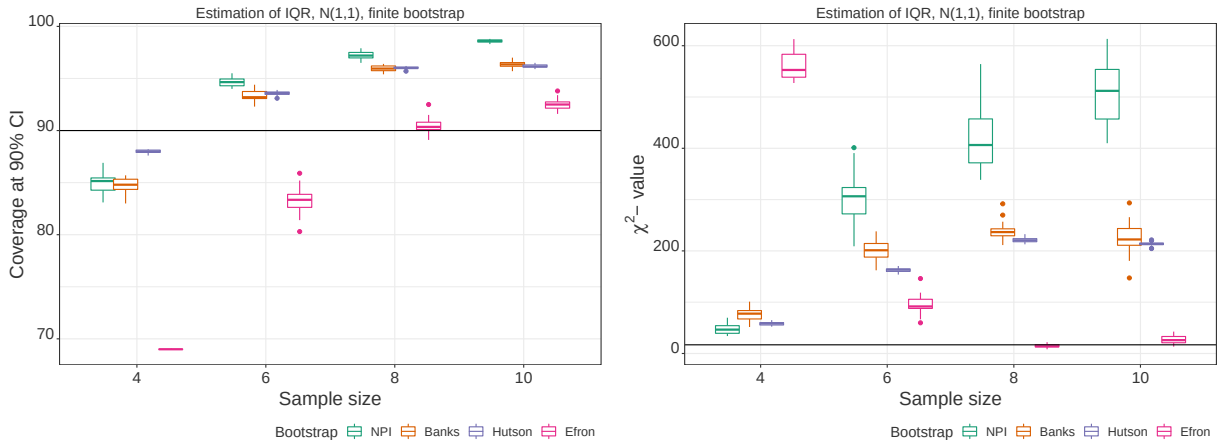


Figure 2.4: Coverage at 90% CI and χ^2-values, estimation of Q1, median and Q3, N(1,1), $n = 4, 6, 8, 10$, finite (Approach I) NPI and Banks-B, 20 simulations



(a) (b)

Figure 2.5: Coverage at 90% CI and chi square values, estimation of IQR, N(1,1), $n = 4, 6, 8, 10$, finite (Approach I) NPI and Banks-B, 20 simulations

perform similarly well. At $n = 10$, Efron-B has the best coverage at 90% CI but it has large χ^2-value, so it does not perform better than Banks-B or Hutson-B. To summarise, simulations for the estimation of Q1, median and Q3 (Figure 2.4) indicate that Banks-B and Hutson-B are suitable choices of the bootstrap method for the estimation of quantiles for Normally distributed small samples ($n = 4, 6, 8, 10$).

Whilst simulations have shown that the smooth bootstrap methods (Banks-B and Hutson-B) perform notably better in the estimation of mean, variance and quantiles for small samples for Normally distributed data, the conclusions for the estimation of IQR (Figure 2.5) are less clear and they differ per size. Both χ^2-value and the coverage at 90% CI increase for Banks-B, NPI-B and Hutson-B as the sample size increases. For $n = 4$, NPI-B has the lowest χ^2-value and similar under-coverage at 90% CI as Banks-B; Hutson-B has the best coverage at 90% CI, only small under-coverage; and Efron-B has large under-coverage. For $n = 6$, Banks-B and Hutson-B have the best coverage (slight over-coverage) and Hutson-B has the second best χ^2-value, and Efron-B has the lowest χ^2-value. From $n = 8$, Efron-B has the lowest χ^2-value and it has the best coverage at 90% CI. Thus, the conclusions about the bootstrap performance in the estimation of IQR for $n = 4, 6$ are unclear, and from $n = 8$, Efron-B would be the preferred choice. Overall, the conclusions regarding the bootstrap methods' performance in the estimation for IQR

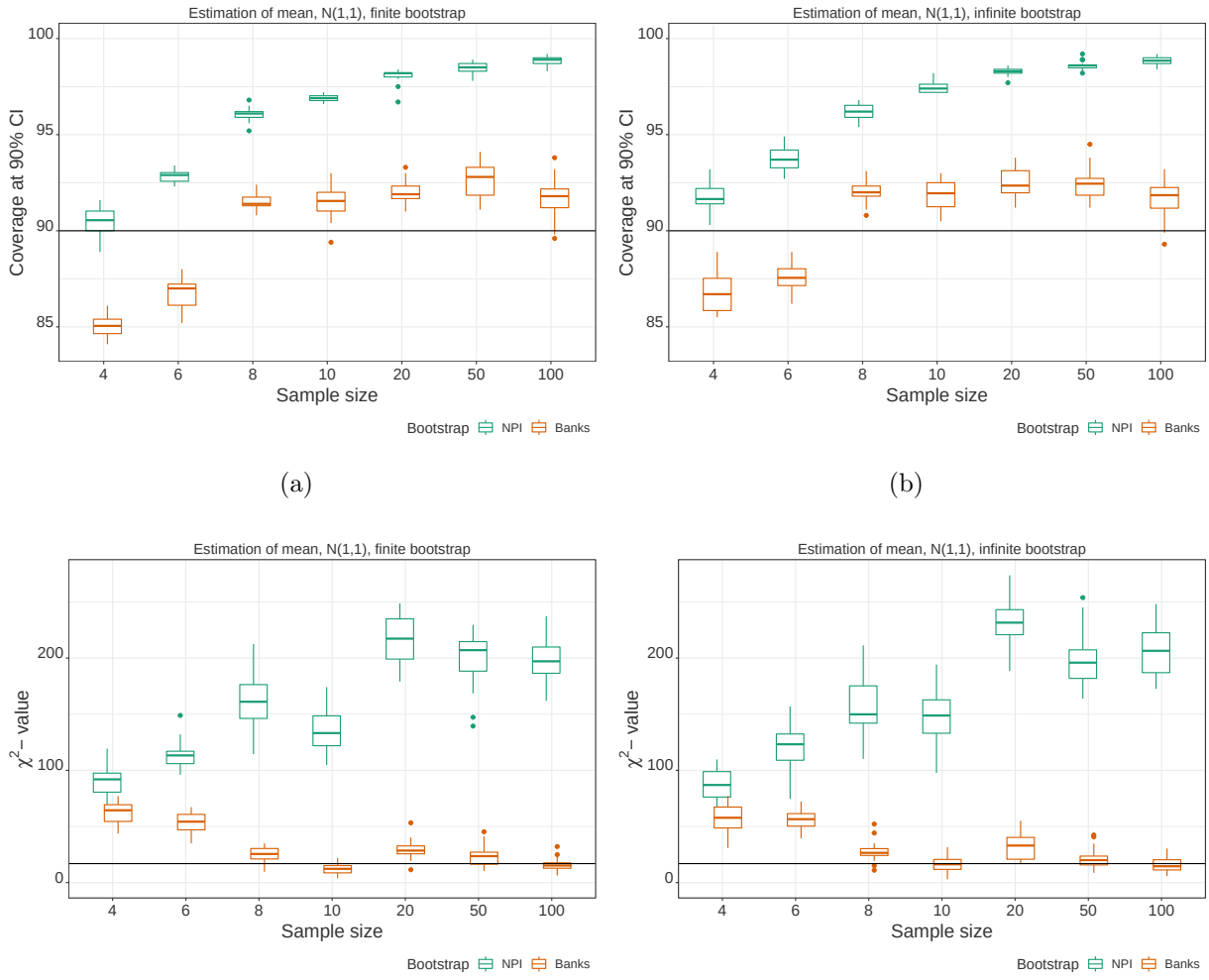


Figure 2.6: Coverage at 90% CI and χ^2-values for NPI-B and Banks-B, estimation of mean, N(1,1), $n = 4, 6, 8, 10, 20, 50, 100$, finite versus infinite range, 20 simulations

for Normally distributed data do not follow a clear pattern, hence a recommendation cannot be given to the researcher.

Choice of range

The performance of NPI-B and Banks-B in the estimation of population characteristics is affected by the choice of range, introduced in Section 2.3.3, for these two bootstrap methods. This effect has been explored for data from Normal distribution. Figures 2.6 and 2.7 display the comparison between using finite (Approach I, Section 2.3.3) and infinite (Approach IV, Section 2.3.3) range for the estimation of mean and variance,

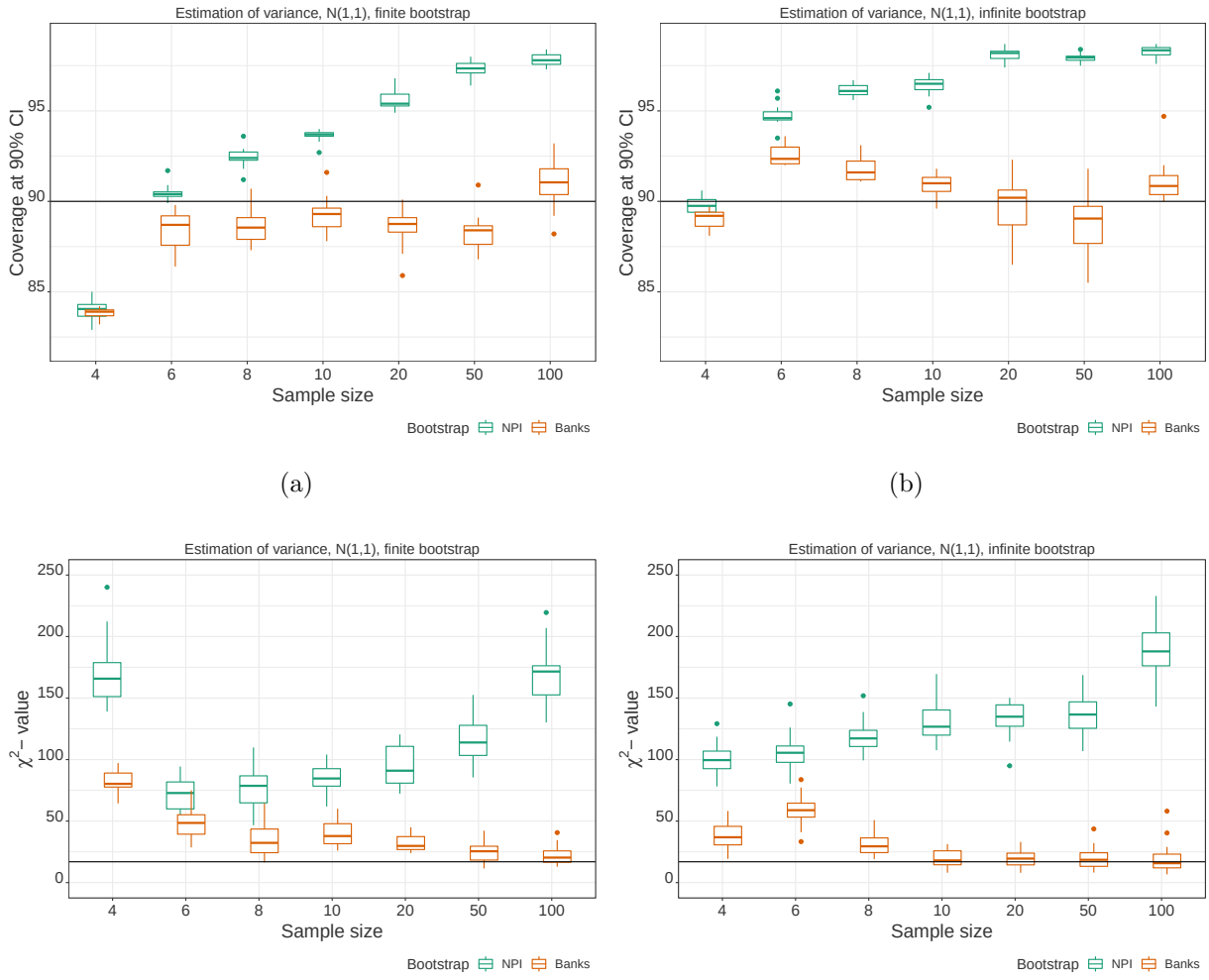


Figure 2.7: Coverage at 90% CI and χ^2-values for NPI-B and Banks-B, estimation of variance, N(1,1), $n = 4, 6, 8, 10, 20, 50, 100$, finite versus infinite range, 20 simulations

respectively. This thesis concludes that for the estimation of mean, median and Q1, the Algorithm 1 outcomes (both coverage at 90% and χ^2-value) are similar for the finite range (Approach I, Section 2.3.3) and for the infinite range (Approach IV, Section 2.3.3). For the estimation of variance, the coverage at 90% CI is better for infinite Banks-B compared to finite Banks-B. For NPI-B, this observation only applies to $n = 4$. An alternative to finite Approach I would be finite Approach III (with $v = 0.1$, Section 2.3.3), the latter approach creates smaller first and last intervals. However, the simulation for the estimation of mean, displayed in Figure 2.8, shows that the use of a smaller first and last interval increases under-coverage at 90% CI (for Banks-B when $n = 4, 6, 8, 10$ and

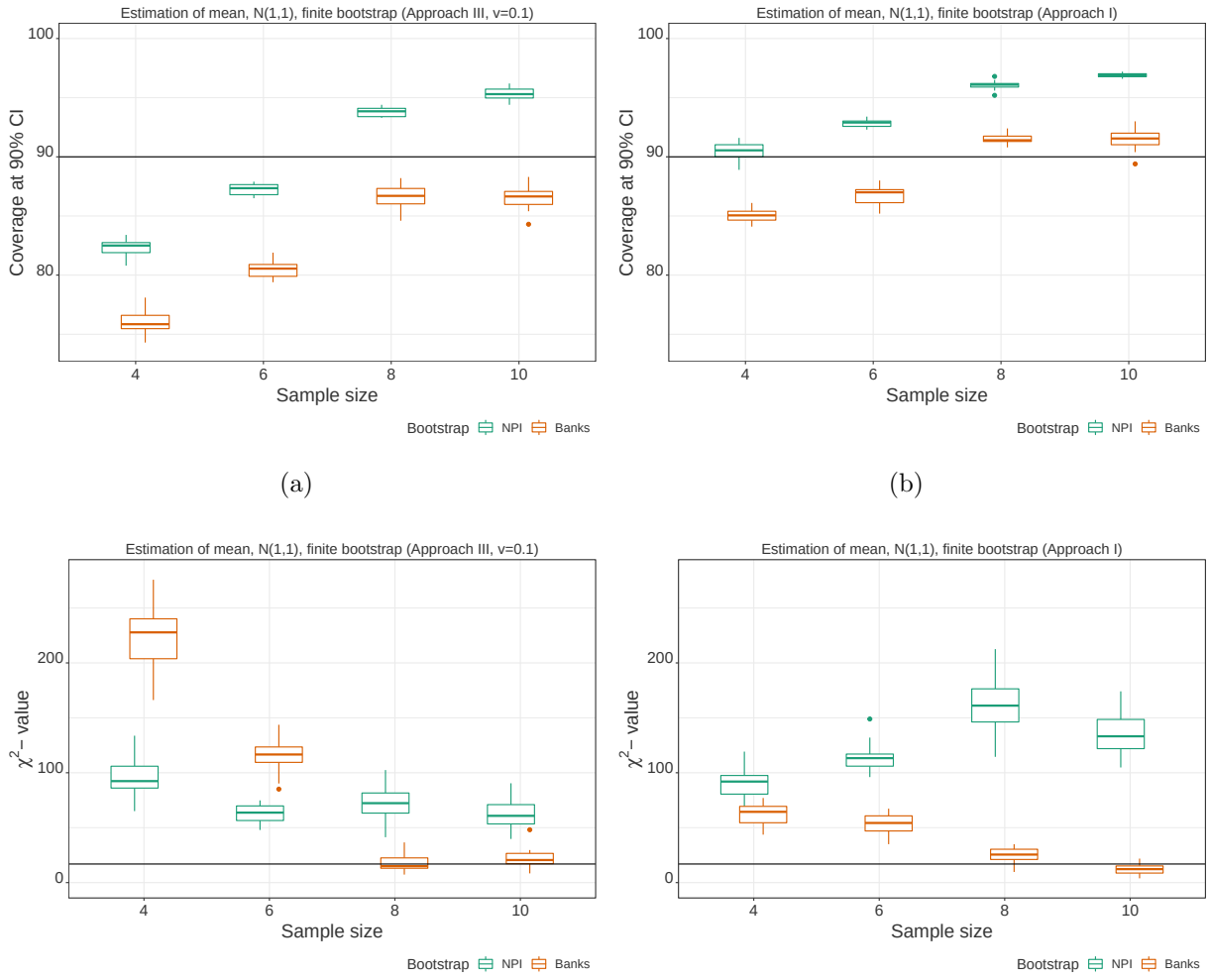


Figure 2.8: Coverage at 90% CI and χ^2-values for NPI-B and Banks-B, estimation of mean, N(1,1), $n = 4, 6, 8, 10$, finite Approach I vs. finite Approach III (with $v = 0.1$), 20 simulations

for NPI-B when $n = 4, 6$) and it leads to higher χ^2-value for Banks-B (when $n = 4, 6$). Given that using finite intervals is computationally simpler, for the estimation of mean and quantiles, this thesis recommends the use of finite range (Approach I, Section 2.3.3) and the use of infinite range (Approach IV, Section 2.3.3) for the estimation of variance. The range is further discussed in Appendix A.4.

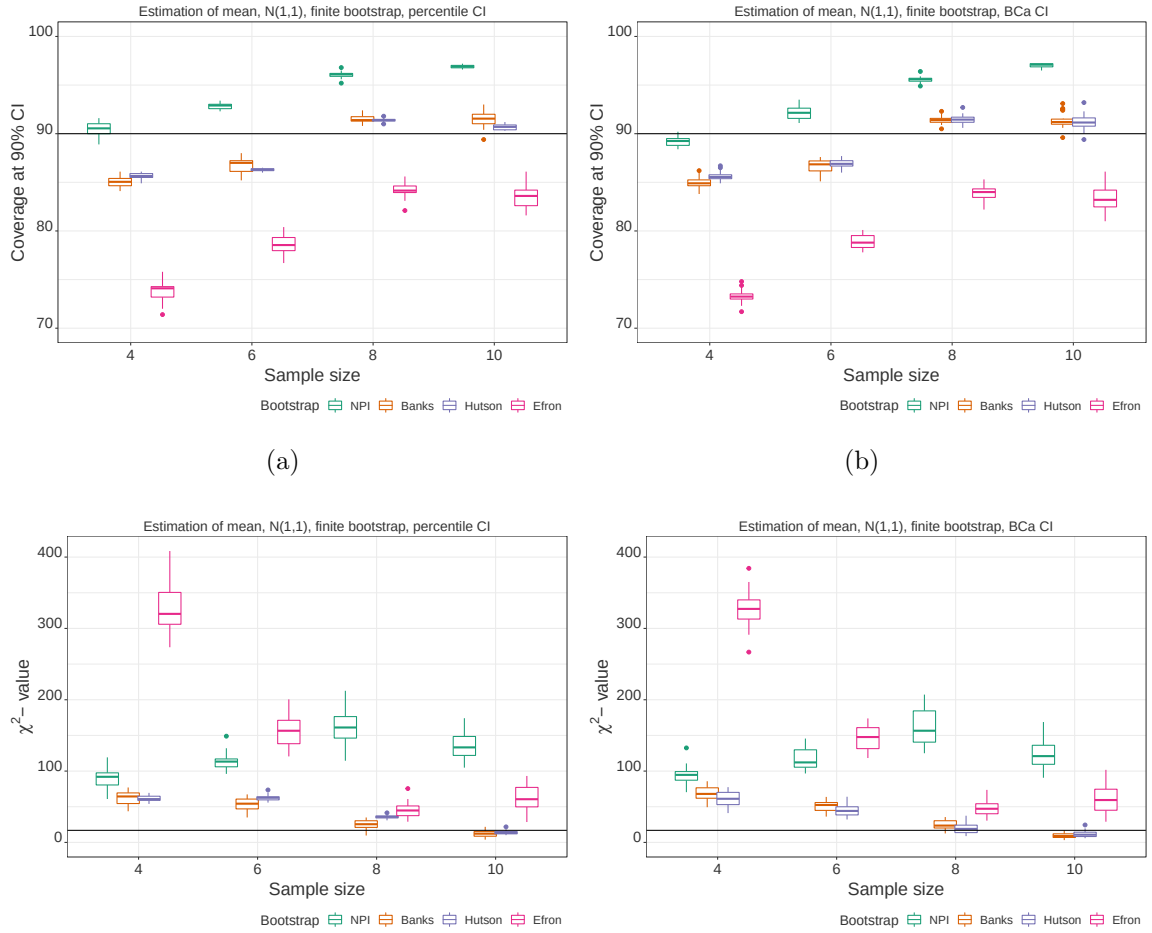


Figure 2.9: Coverage at 90% CI and χ^2-values, estimation of mean, N(1,1), $n = 4, 6, 8, 10$, finite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

BC_a confidence intervals for Normally distributed data

So far, percentile confidence intervals were used. This investigation also explored how would the use of $(1-2\alpha)\%$ BC_a confidence intervals instead of $(1-2\alpha)\%$ percentile confidence interval affect the performance of the four bootstrap methods. In Algorithm 1, BC_a confidence intervals were calculated. The method for calculating BC_a confidence intervals can be found in Section 2.3.5. The comparisons between percentile versus BC_a confidence intervals in the estimation of mean, variance, median, Q1, Q3 and IQR are illustrated in Figures 2.8-2.13. The observations will be discussed separately for each population characteristic.

In the estimation of mean (Figure 2.9), the simulation outputs for BC_a confidence

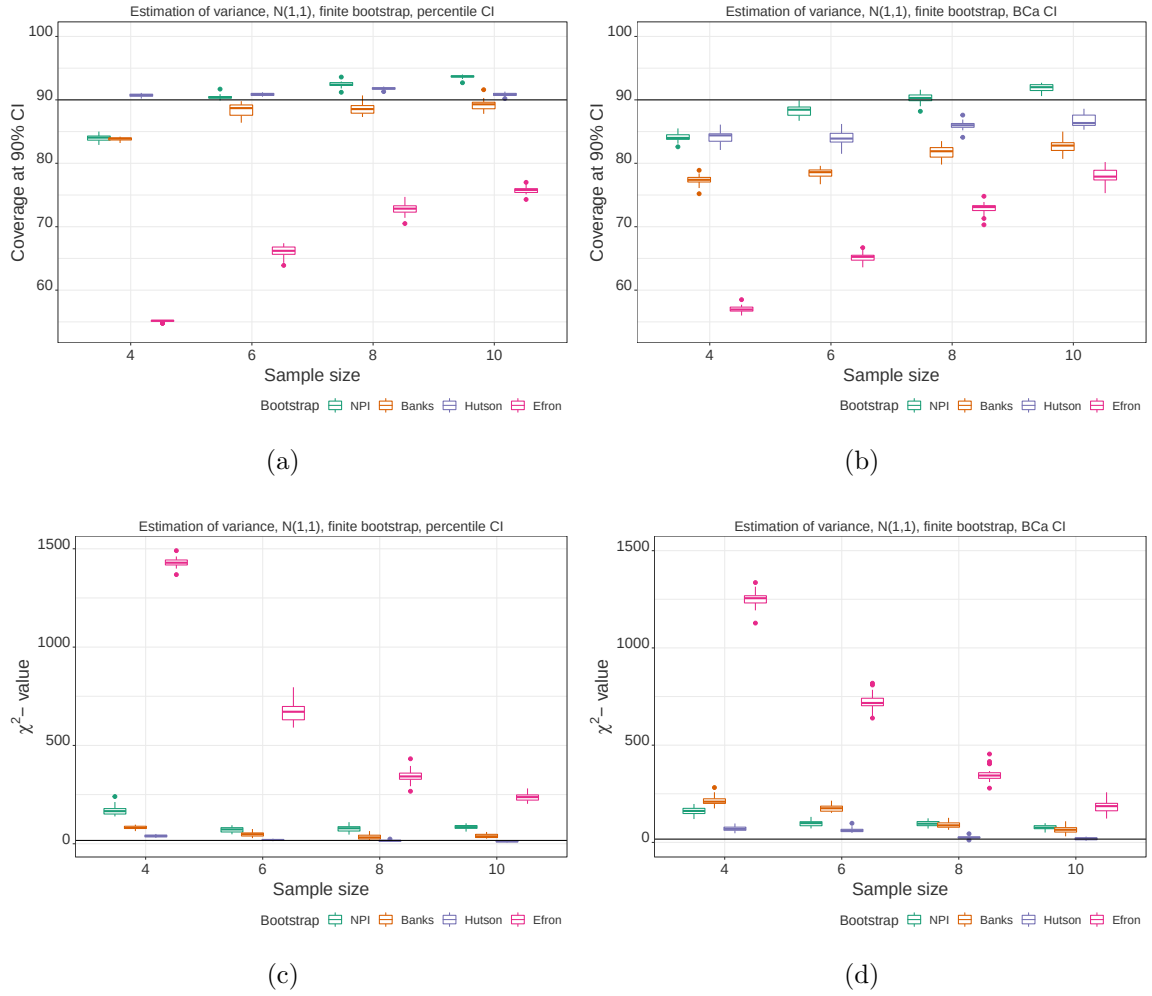


Figure 2.10: Coverage at 90% CI and χ^2-values, estimation of variance, N(1,1), $n = 4, 6, 8, 10$, finite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

intervals are similar to outputs for percentile confidence intervals for NPI-B, Banks-B and Efron-B. The biggest difference is apparent for Hutson-B. For the estimation of mean, Hutson-B with BC_a confidence intervals has lower χ^2-value for $n = 4, 6, 8$ than Hutson-B with percentile confidence intervals, and, as a consequence, Hutson-B slightly outperforms Banks-B for these sample sizes.

By contrast, in the estimation of variance, Hutson-B with BC_a confidence intervals performs worse than with percentile confidence intervals, as the use of BC_a confidence intervals leads to lower χ^2-value for all n. Similarly, Banks-B performs worse in the estimation of variance when BC_a confidence intervals are used instead of percentile confidence

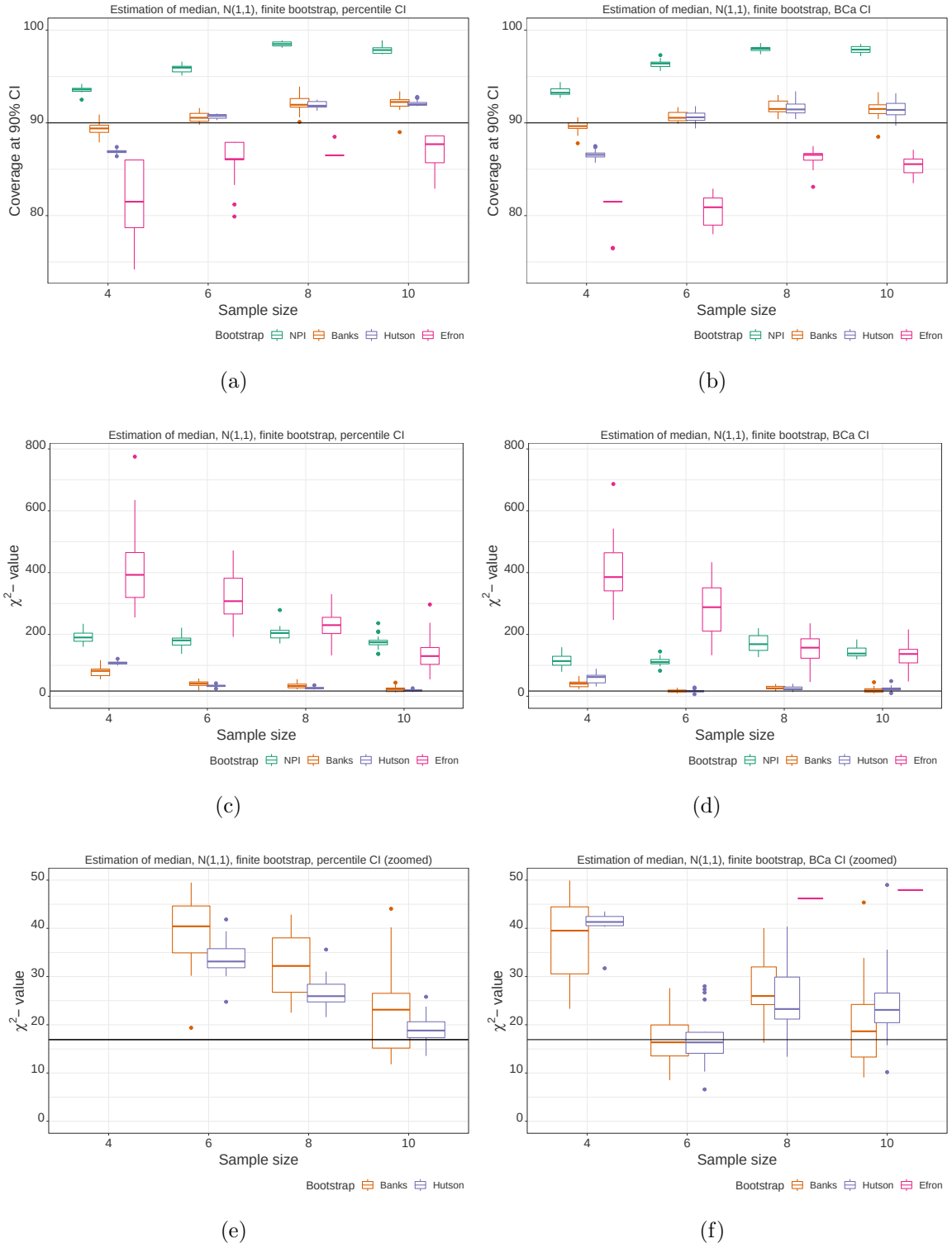


Figure 2.11: Coverage at 90% CI and χ^2-values, estimation of median, N(1,1), $n = 4, 6, 8, 10$, finite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

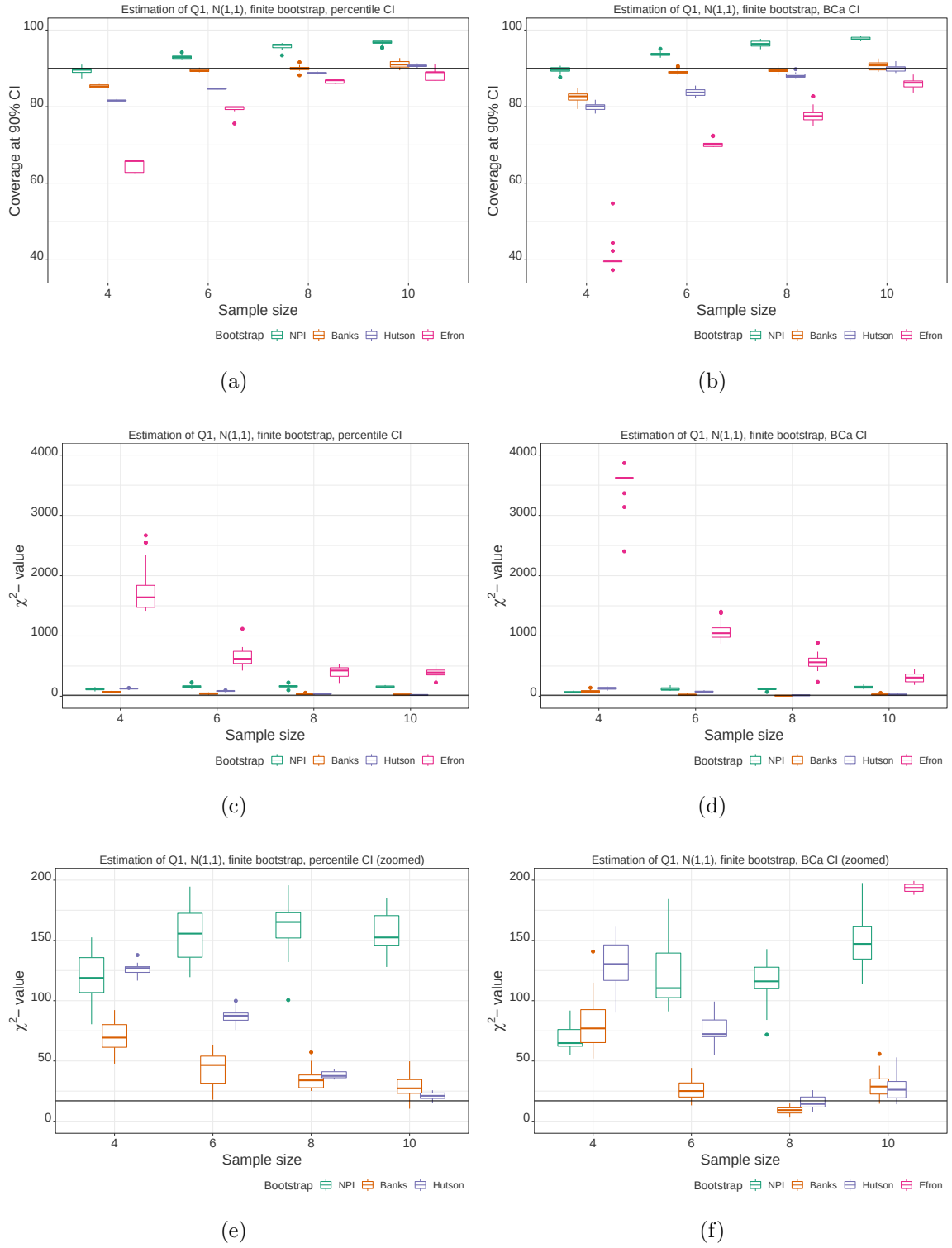


Figure 2.12: Coverage at 90% CI and χ^2-values, estimation of Q1, N(1,1), $n = 4, 6, 8, 10$, finite NPI and Banks-B, percentile versus BC$_a$ confidence intervals, 20 simulations

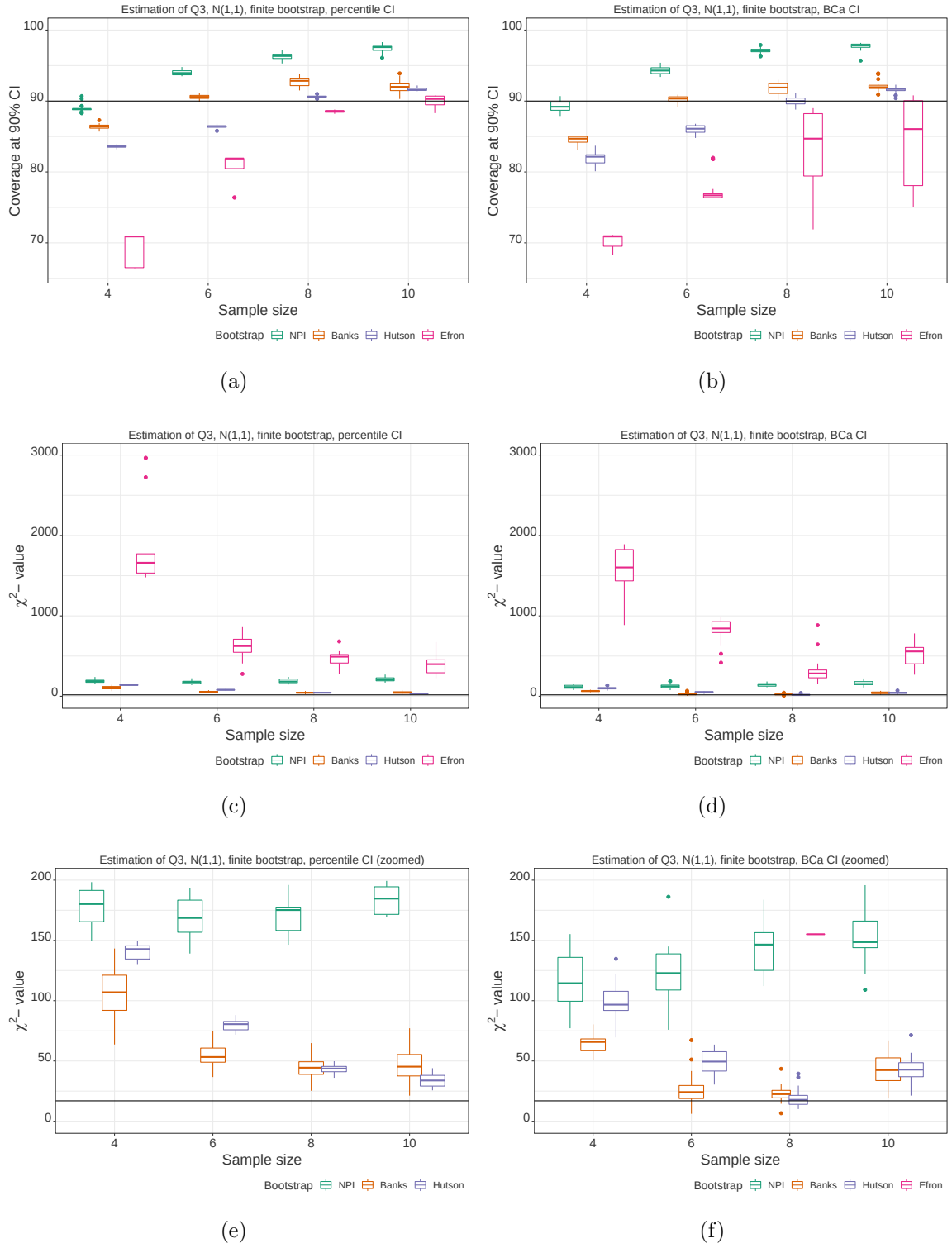


Figure 2.13: Coverage at 90% CI and χ^2-values, estimation of Q3, N(1,1), $n = 4, 6, 8, 10$, finite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

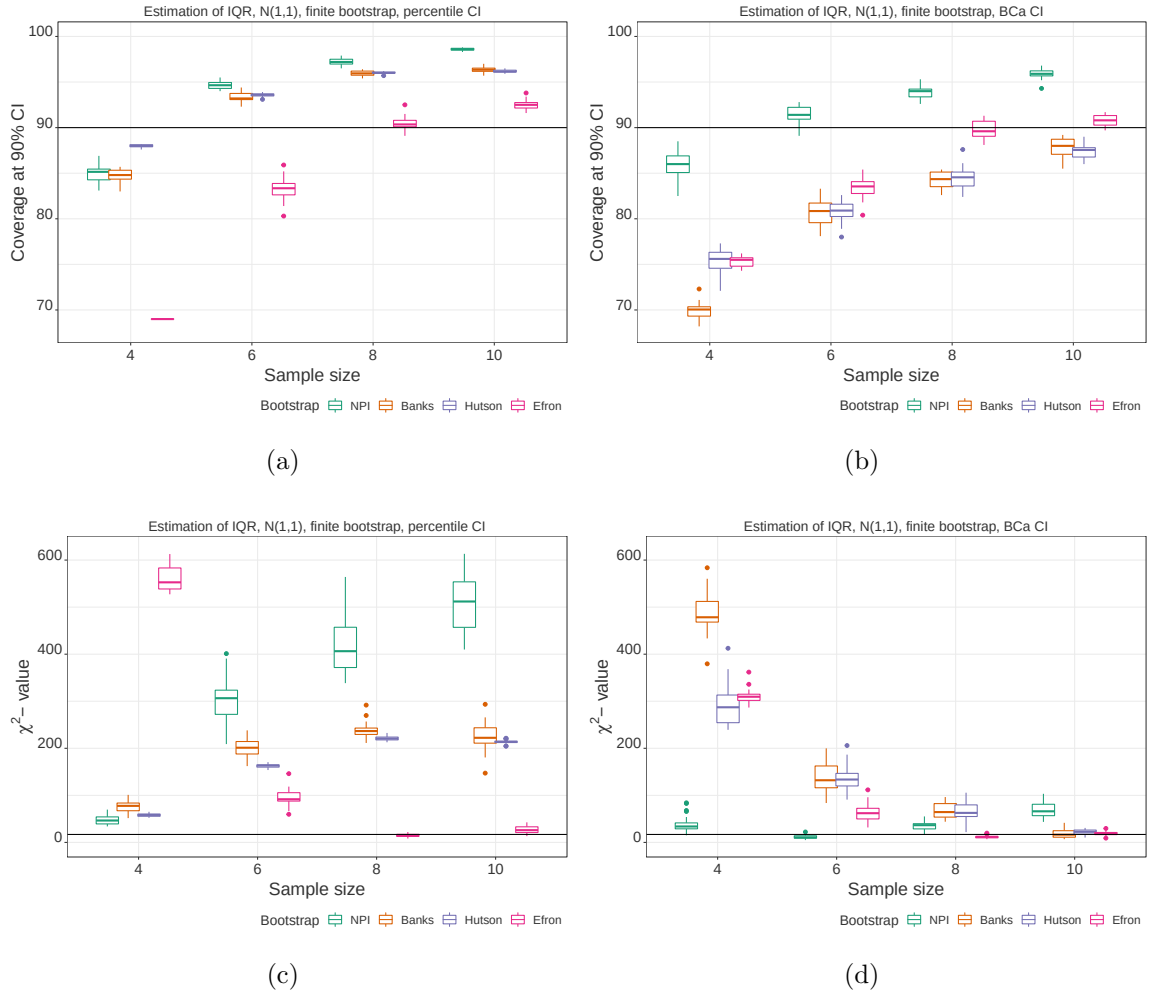


Figure 2.14: Coverage at 90% CI and χ^2-values, estimation of IQR, N(1,1), $n = 4, 6, 8, 10$, finite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

intervals. Efron-B with BC_a confidence intervals performs slightly better for $n = 4, 10$, but it still remains the worst performing bootstrap method in the estimation of variance for Normally distributed data with sample size $n \leq 10$. NPI-B is the least affected bootstrap method by the choice of confidence intervals.

In the estimation of median (Figure 2.11), Efron-B with BC_a confidence intervals does not perform better than Banks-B or Hutson-B for small samples, but it outperforms NPI-B for $n = 8, 10$. Both Banks-B and NPI-B with BC_a confidence intervals have lower χ^2-values for all studied sample sizes than the respective bootstrap methods with percentile confidence intervals. Hutson-B with BC_a confidence intervals has slightly lower χ^2-value for $n = 4, 6, 8$ than Hutson-B with percentile confidence intervals.

In the estimation of Q1 (Figure 2.12), Efron-B with BC_a confidence intervals performs worse than Efron-B with percentile confidence intervals: it has worse χ^2-values for $n = 4, 6, 8$ and it has worse coverage at 90% CI. NPI-B and Banks-B with BC_a confidence intervals have lower χ^2-values for $n = 4, 6, 8$ than those bootstrap methods with percentile confidence intervals and Hutson-B performs better with BC_a confidence intervals than with percentile confidence intervals for $n = 4, 6$ (by assessing the χ^2-value).

In the estimation of Q3 (Figure 2.13), Banks-B and Hutson-B have lower χ^2-value for $n = 4, 6, 8$ when BC_a confidence intervals are used and NPI-B with BC_a confidence intervals has lower χ^2-value than NPI-B with percentile confidence intervals for all sample sizes. Efron-B with BC_a confidence intervals has worse coverage at 90% CI than Efron-B with percentile confidence intervals for $n = 6, 8, 10$.

Lastly, we do not recommend using BC_a confidence intervals for the estimation of IQR for Banks-B and Hutson-B, as it makes the coverage worse (see Figure 2.14). On the other hand, NPI-B with BC_a confidence intervals performs notably better in the estimation of IQR than NPI-B with percentile confidence intervals for all n, considering both metrics of assessment. Efron-B with BC_a confidence intervals performs better than Efron-B with percentile confidence intervals at $n = 4$ from the perspective of both metrics and it has lower χ^2-value for $n = 6, 8, 10$. Thus, we conclude that BC_a confidence intervals have biggest effect on the estimation of IQR for small sample sizes and their use could be considered for Efron-B and NPI-B in the estimation of IQR. It can be concluded that NPI-B with BC_a confidence intervals performs the best in the estimation of IQR for $n = 4, 6$, and Efron-B with BC_a confidence intervals performs the best in the estimation of IQR for $n = 8, 10$.

In summary, using BC_a confidence intervals instead of percentile confidence intervals does not make Efron-B a better performing bootstrap method, in the estimation of mean, variance and quantiles for small samples, than the other three bootstrap methods (Banks-B, Hutson-B and NPI-B) from the perspective of either of the two metrics of assessment. We would recommend the use of BC_a confidence for Efron-B only for the estimation of IQR. For further application of BC_a CI with Banks-B, Hutson-B and NPI-B for the estimation of population characteristics of Normally distributed data of small sample size, it is essential to carry out further research into the performance of BC_a confidence

intervals for Hutson-B for the estimation of mean and quantiles (Q1, median and Q3) and for Banks-B and NPI-B for the estimation of quantiles (Q1, median and Q3).

Further investigation topics

This section has shown that adjusting variables, such as the type of confidence intervals or the type of range for NPI-B or Banks-B, has an effect on the bootstrap method's performance in estimation. Some initial results of two further topics are discussed in Appendices A.2 and A.3. A brief investigation into how the bootstrap methods perform in estimation when the sample size is as small as $n = 3$, or even smaller, $n = 2$, has been carried out and it can be concluded that Banks-B, Hutson-B and NPI-B can be still considered for the use for the estimation of mean, variance and quantiles for Normally distributed data, however, the outcomes of analysis should be considered with great care, given that there is no guarantee that such small sample is an accurate representation of the population. Further details are reported in Appendix A.2. Moreover, initial investigation into the variability of the bootstrap methods outcomes for Normally distributed data is discussed Appendix A.3.

2.4.3 Lognormally and Exponentially distributed data

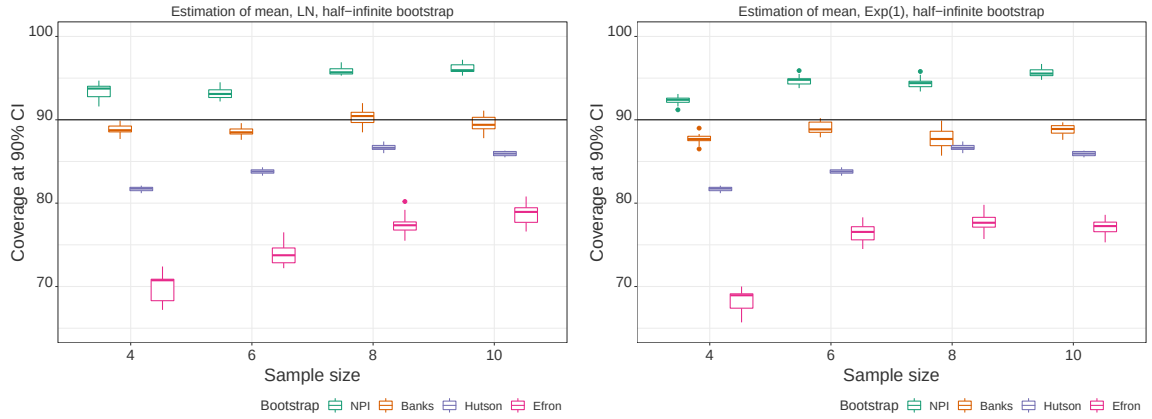
This section presents findings on the bootstrap performance in the estimation of population characteristics for data from Lognormal and Exponential distribution. Lognormal distribution is an example of a skewed distribution, with a heavy tail. Lognormal distribution has probability density function $f(x) = \frac{1}{xs_{LN}\sqrt{2\pi}} e^{-\frac{(\ln x - m_{LN})^2}{2s_{LN}^2}}$ for $x > 0$, $s_{LN} > 0$, $m_{LN} \in (-\infty, \infty)$. The distribution parameters of the Lognormal distribution, m_{LN} and s_{LN}, are calculated from the mean μ and variance σ^2 of the Normal distribution using the following formulas: $m_{LN} = \ln(\mu^2/\sqrt{\mu^2 + \sigma^2})$ and $s_{LN} = \sqrt{\ln(\sigma^2/(\mu^2 + 1))}$. For Lognormal distribution, the chosen mean and standard deviation are $\mu = 1$ and $\sigma = 1$, respectively, which leads to Lognormal distribution LN($m_{LN} = -0.3465736$,$s_{LN}^2 = 0.8325546^2$). When we refer to Lognormal distribution later, we round these distribution parameters to three decimal places. For Lognormal distribution, the true population characteristics calculations are: median for LN(m_{LN}, s_{LN}^2) is equal to $e^{m_{LN}+s_{LN}*\Phi^{-1}(0.5)}$, Q1 for LN($m_{LN}, s_{LN}^2$) equals to $e^{m_{LN}+s_{LN}*\Phi^{-1}(0.25)}$ and Q3 for LN(m_{LN}, s_{LN}^2) is equal

to $e^{m_{LN}+s_{LN}*\Phi^{-1}(0.75)}$. Φ stands for the cumulative distribution function of the standard Normal distribution. Thus, true Q1 for LN($m_{LN} = -0.347$,$s^2_{LN} = 0.833^2$) is 0.40328, true median for LN($m_{LN} = -0.347$,$s^2_{LN} = 0.833^2$) is 0.70711 and true Q3 for LN($m_{LN} = -0.347$,$s^2_{LN} = 0.833^2$) is 1.23983.

Exponential distribution has probability density function $f(x;\lambda) = \lambda e^{\lambda x}$ for $x \geq 0$. The data are generated from Exponential distribution with rate $\lambda = 1$. For Exponential distribution, the true parameter calculations are: mean for Exp(λ) is equal to $1/\lambda$, median for Exp(λ) is equal to $\ln(2)/\lambda$, variance for Exp(λ) is equal to $1/\lambda^2$, Q1 for Exp(λ) is equal to $\ln(4/3)/\lambda$, Q3 for Exp(λ) is equal to $\ln(4)/\lambda$ and IQR for Exp(λ) is equal to $\ln(4)/\lambda - \ln(4/3)/\lambda$. Thus, true mean for Exp(1) is 1, true median for Exp(1) is 0.69315, true variance for Exp(1) is 1, true Q1 for Exp(1) is 0.28768, true Q3 for Exp(1) is 1.38629 and true IQR for Exp(1) is 1.09861.

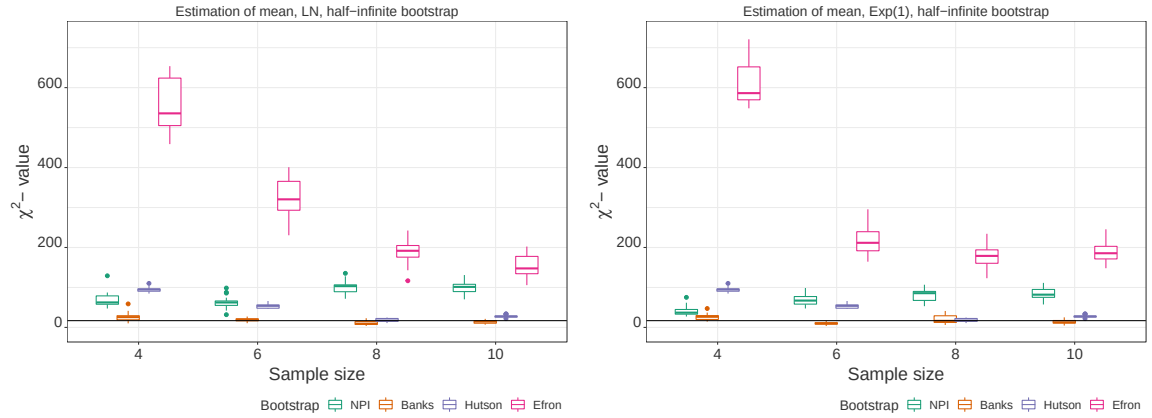
Estimation of mean

Both Exponential and Lognormal distributions are only defined on $[0, \infty)$. Thus, for Banks-B, NPI-B and Hutson-B, the half-infinite range, $[0, \infty)$, is first assumed: for Banks-B and NPI-B, Approach V (see Section 2.3.3) is employed and for Hutson-B Equation (2.3) from Section 2.3.4 is used. Comments on the effect of the choice of tails will be provided later. The discussion begins with the bootstrap methods performance in the estimation of mean for data from Exponential and Lognormal distribution (Figure 2.15), when percentile confidence intervals are calculated. The performance of Efron-B in the estimation of mean for data from Lognormal distribution and Exponential distribution is very poor (high χ^2-value and large under-coverage at 90% CI). Banks-B, Hutson-B and NPI-B perform better in the estimation of mean. Coverage at 90% CI for Banks-B is close to 900 (the ideal coverage), there is a small under-coverage for Exponentially distributed data for all sample sizes and for Lognormally distributed data for $n = 4, 6$. For NPI-B, there is over-coverage for all sample sizes. Hutson-B has under-coverage at 90% CI for both distributions and for all sample sizes. Considering the second metric, χ^2-value, it can be concluded that half-infinite Banks-B has the lowest χ^2-value for all the presented sample sizes for both Exponential and Lognormal distribution, therefore, it is the best performing bootstrap method in the estimation of mean. Hutson-B performs worse in



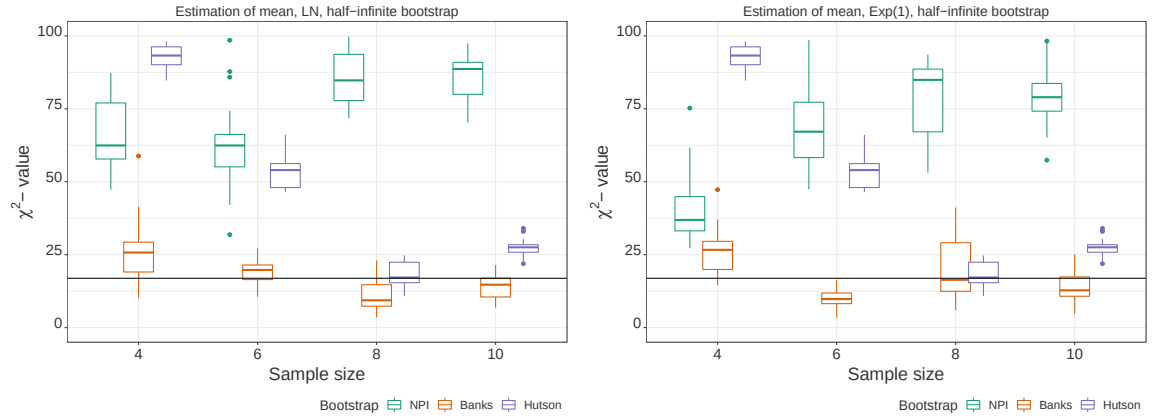
(a)

(b)



(c)

(d)



(e)

(f)

Figure 2.15: Coverage at 90% CI and χ^2-values, estimation of mean, LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, 20 simulations

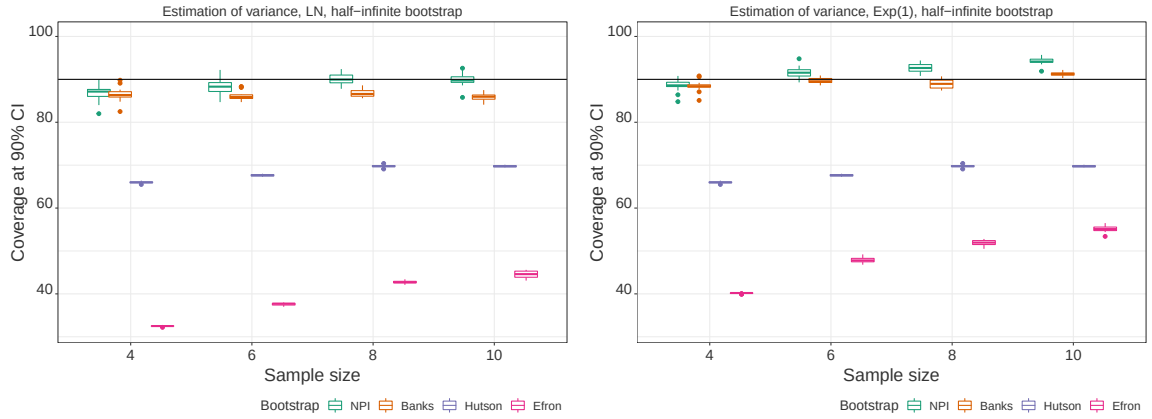
the estimation of mean for data that are not Normally distributed than for Normally distributed data.

Estimation of variance

Conclusions for the estimation of variance (Figure 2.16) are different for data from Exponential and Lognormal distributions. For Exponentially distributed data, Banks-B is the best performing bootstrap in the estimation of variance as it has the lowest χ^2-value for $n = 4, 6, 8, 10$. On the other hand, for Lognormally distributed data, NPI-B is the best performing bootstrap in the estimation of variance as it has the best coverage at 90% CI and the lowest χ^2-value for small sample sizes, $n = 4, 6, 8, 10$. The latter observations could be linked to the fact that Lognormal distribution is a skewed distribution. Arguably, NPI-B is better at capturing the non-symmetric shape of the distribution. Efron-B performs poorly in the estimation of variance for Exponentially and Lognormally distributed data (large under-coverage at 90% CI and high χ^2-value). Hutson-B is the third best performing bootstrap method, but it still performs poorly from the perspective of both metrics of assessment. In Section 2.4.2 it was shown that Hutson-B performs well in the estimation of variance for small samples with Normal underlying distribution. However, this conclusion does not extend to Exponentially and Lognormally distributed data. Thus, the reader should be careful about using Hutson-B for the estimation of variance, given that one cannot ascertain Normal distribution for small samples.

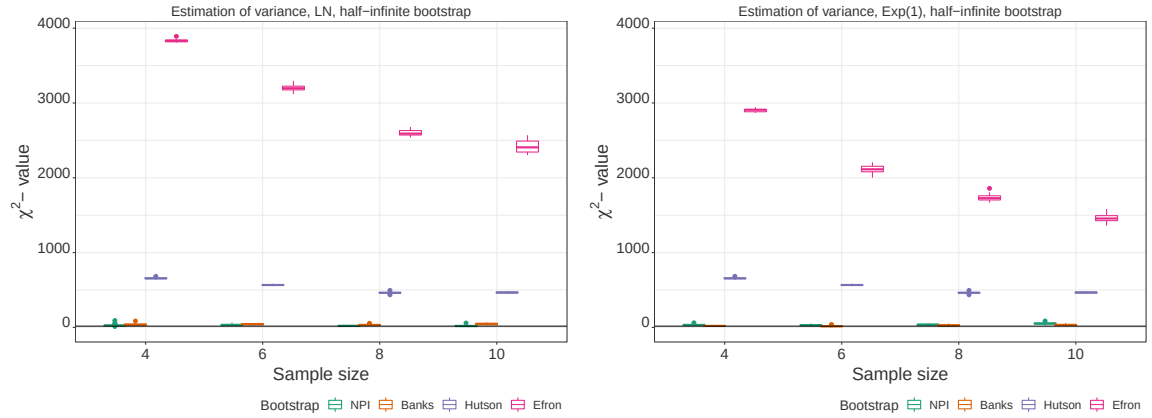
Estimation of quantiles and IQR

The bootstrap performance in the estimation of quantiles and IQR has been explored for Exponentially and Lognormally distributed data (Figures 2.17, 2.18, 2.19 and 2.20). Banks-B is a suitable bootstrap method for the estimation of quantiles (Q1, median and Q3) for both Exponentially and Lognormally distributed data as it has good coverage and low χ^2-value. NPI performs well in the estimation of Q3 for $n = 4$ for Exponentially distributed data. For Lognormally distributed data, NPI-B also performs well at $n = 4$ but Banks-B is clearly still the best performing bootstrap method here. Hutson-B also performs well, in some cases better than Banks-B. Observations for different quantiles will be discussed separately.



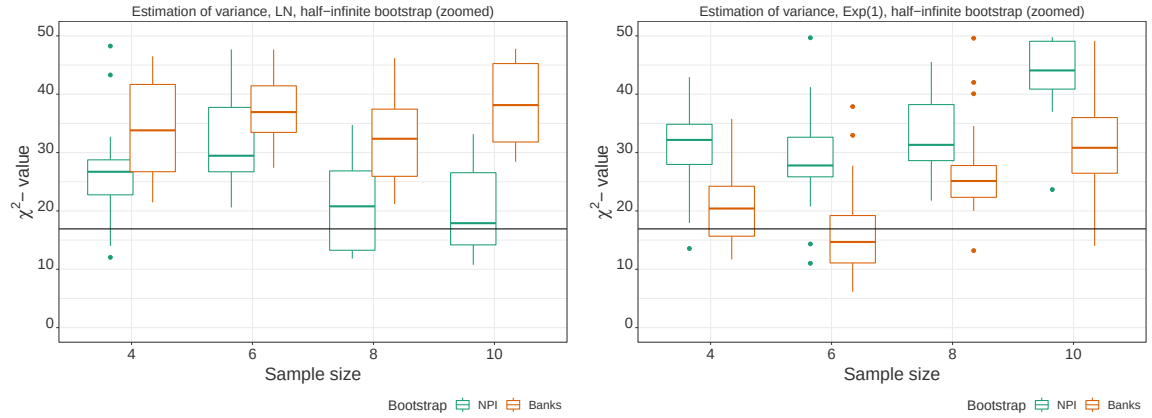
(a)

(b)



(c)

(d)



(e)

(f)

Figure 2.16: Coverage at 90% CI and χ^2-values, estimation of variance, LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, 20 simulations

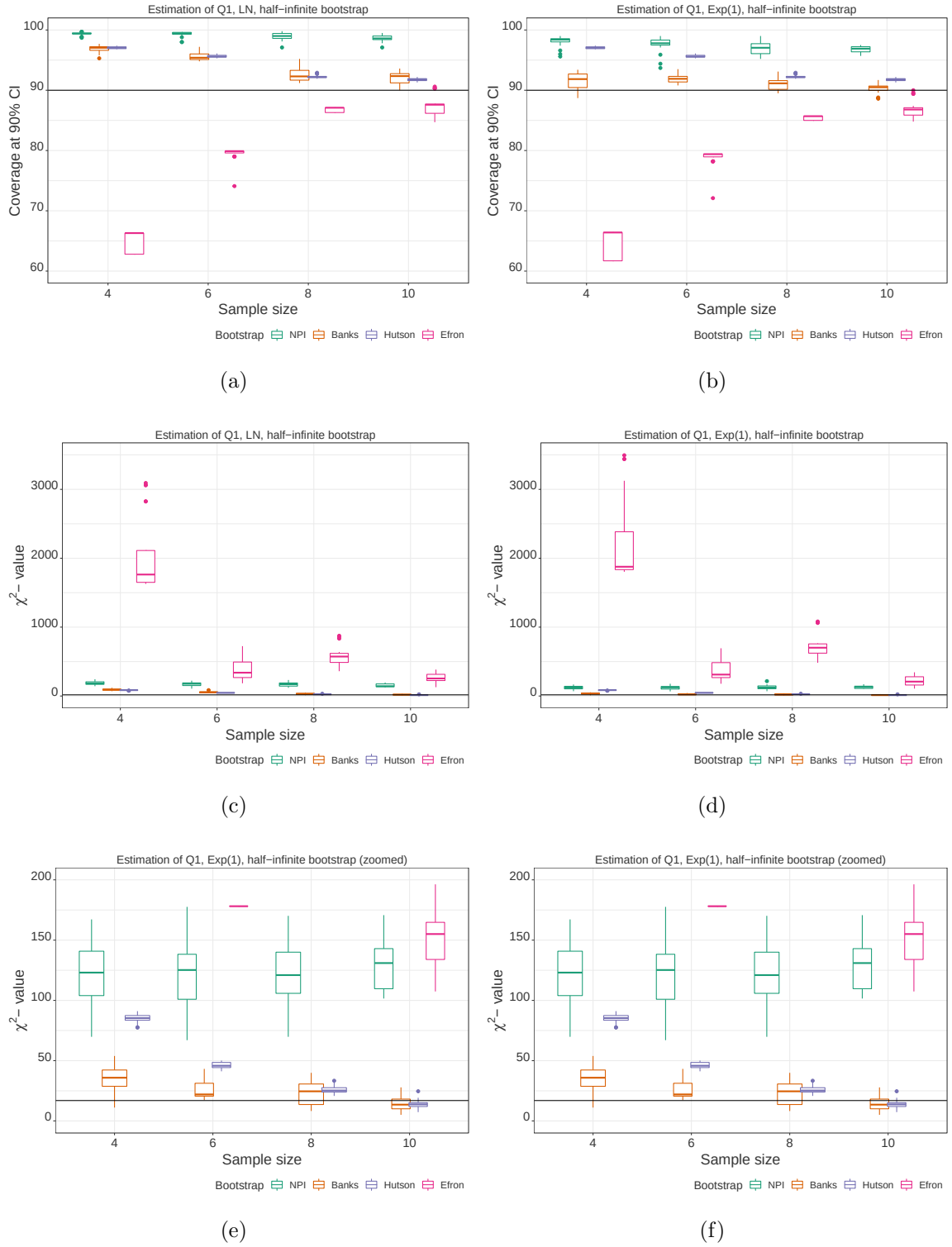


Figure 2.17: Coverage at 90% CI and χ^2-values, estimation of Q1, LN ($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, 20 simulations

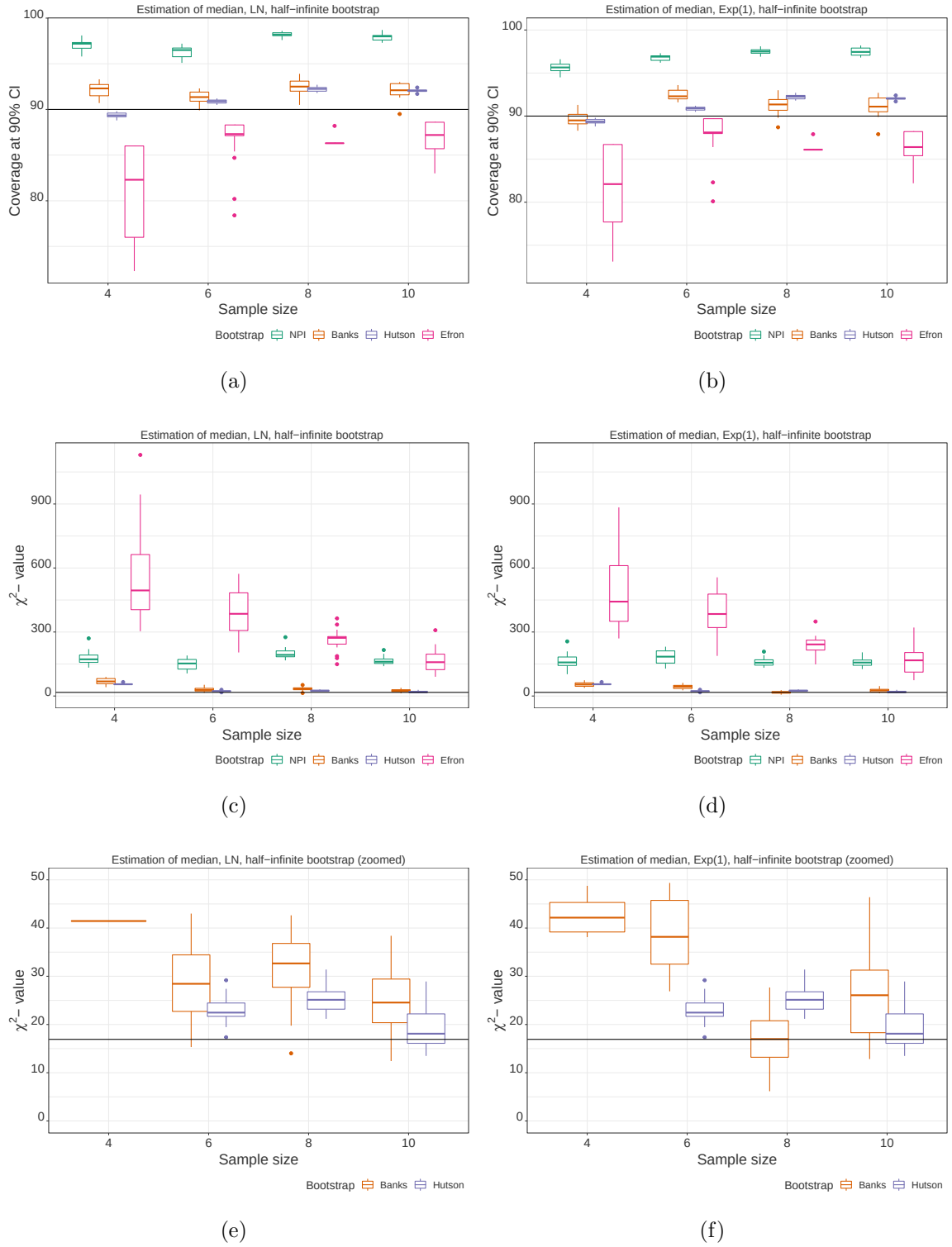


Figure 2.18: Coverage at 90% CI and χ^2-values, estimation of median, LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, 20 simulations

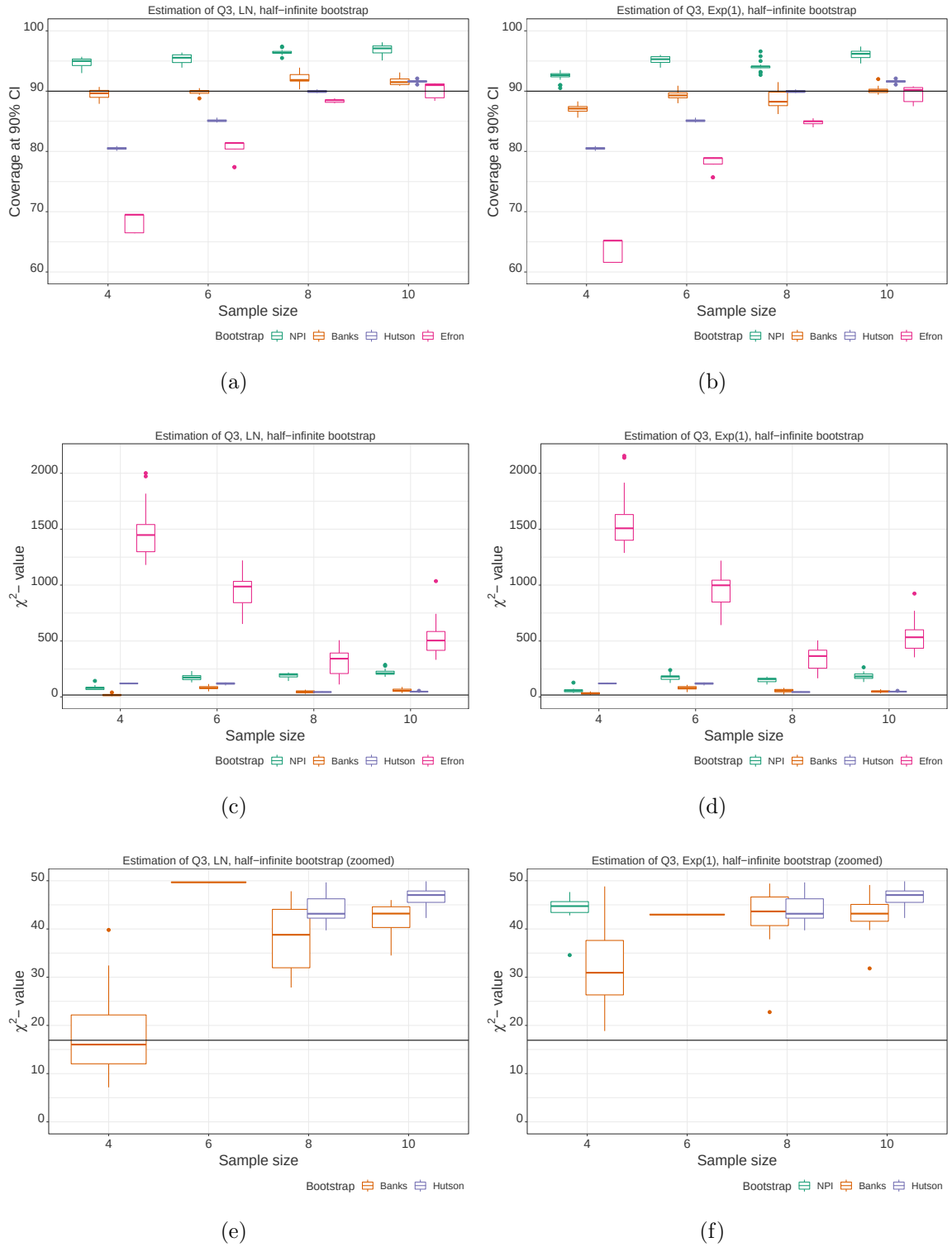


Figure 2.19: Coverage at 90% CI and χ^2-values, estimation of Q3, LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, half-infinite bootstrap NPI and Banks-B, 20 simulations

In the estimation of Q1 (Figure 2.17), Banks-B is the best performing bootstrap (lowest χ^2-value for $n = 4, 6, 8$ and small over-coverage – close to the ideal coverage) for Exponentially distributed data. For Lognormally distributed data, Banks-B and Hutson-B both perform well in the estimation of Q1: they have similar coverage at 90% CI (smaller over-coverage than NPI-B, close to the ideal coverage), but Hutson-B has the lowest χ^2-value.

In the estimation of median (Figure 2.18) for both Exponentially and Lognormally distributed data, Banks-B performs the best in the estimation of median for $n = 4$. For Lognormally distributed data, Hutson-B performs slightly better than Banks-B (lower χ^2-value) for $n = 6, 8, 10$ and for Exponentially distributed data, Hutson-B performs better for $n = 6, 10$ and Banks-B is the best performing bootstrap method for $n = 8$. Overall conclusion is that both Banks-B and Hutson-B are good in the estimation of median for both Exponentially and Lognormally distributed data.

In the estimation of Q3 (Figure 2.19) for Exponentially distributed data, for $n = 4$, NPI-B performs well (small over-coverage as opposed to other bootstrap methods that have under-coverage, and second lowest χ^2-value), and Banks-B has under-coverage, but it has the lowest χ^2-value. For $n = 6$, Banks-B has only small under-coverage and it has the lowest χ^2-value. For $n = 8$, Hutson-B has better coverage at 90% CI and slightly lower χ^2-value. For $n = 10$, Banks-B has coverage closest to the ideal coverage and it has the lowest χ^2-value. Thus, for Exponentially distributed data, Banks-B is a suitable recommended bootstrap method. For $n = 8, 10$, Hutson-B also performs well. For Lognormally distributed data, Banks-B is the best performing bootstrap method in the estimation of Q3, as it has good coverage at 90% CI and the lowest χ^2-value. Hutson-B performs worse for Lognormally distributed data, this could be caused by the influence of heavy tails, for which Hutson-B is not an ideal method, as discussed in Section 2.3.4. Overall, for both Exponential and Lognormal distribution, the recommendation stemming out of the simulation study is to use Banks-B for the estimation of quantiles (Q1, median, and Q3) and for some cases also Hutson-B.

It is less clear which bootstrap method performs the best in the estimation of IQR (Figure 2.20). For Exponentially distributed data, at $n = 4$, NPI-B has the best coverage at 90% CI and the lowest χ^2-value, and at $n = 6$, Banks-B has good coverage at 90%

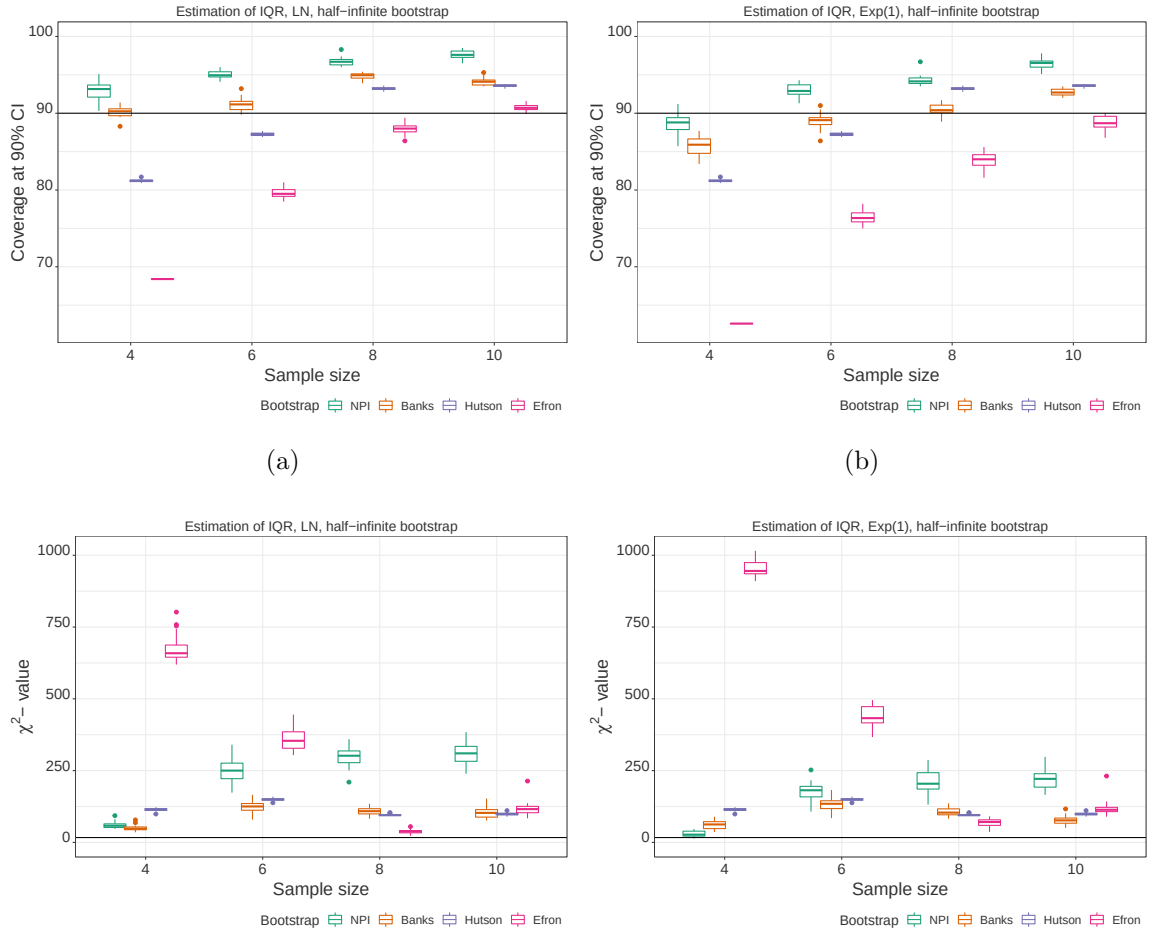


Figure 2.20: Coverage at 90% CI and χ^2-values, estimation of IQR, LN ($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, 20 simulations

CI and the lowest χ^2-value. However, at $n = 8$, Efron-B has the lowest χ^2-values whilst Banks-B has the best coverage at 90% CI, so it is not clear which bootstrap method is the best. Arguably, Banks-B performs the best at $n = 8$ because under-coverage is bad. For $n = 10$, Banks-B has the lowest χ^2-value and there is only small over-coverage, so arguably it is the best performing bootstrap method at this sample size. For Lognormally distributed data, Banks-B is the best performing bootstrap method for $n = 4, 6$ (best coverage and smallest χ^2-value). At $n = 8$, Efron-B has the lowest χ^2-value, but it has a slight under-coverage, so arguable Banks-B and Hutson-B still perform better as they both have over-coverage, rather than under-coverage. At $n = 10$, Efron-B has the best coverage, but Hutson-B has the lowest χ^2-value, therefore, conclusions are not clear.

Before giving more guidance on what bootstrap method to choose for the estimation of IQR for Exponentially and Lognormally distributed data, more exploration needs to be carried out.

Choice of range

So far, half-infinite range (Approach V, Section 2.3.3) was assumed for NPI-B and Banks-B in the simulation study for Exponential and Lognormal distributions. The same simulations were carried out using finite range (Approach I, Section 2.3.3). The coverage at 90% CI is better and the χ^2-value is lower for Banks-B and NPI-B with half-infinite range than for the finite range for small n. For more detail, see Appendix A.4.2. Thus, half-infinite range is recommended for Exponentially and Lognormally distributed data defined on $[0, \infty)$. In practice, it is not always clear whether the data comes from distribution defined on real-line or on on $[0, \infty)$. A researcher could use biological knowledge and use half-infinite range if the data has to be equal or greater than 0.

BC_a confidence intervals for Lognormally distributed data

Similarly to the simulation study for Normally distributed data, the effect of using BC_a instead of percentile confidence intervals in Algorithm 1 for data from Lognormal distribution has been studied. Using BC_a confidence intervals does not affect the previously described findings: Banks-B still performs the best in the estimation of mean and quantiles, and NPI-B performs the best in the estimation of variance for data with Lognormal underlying distribution.

Further detailed observations follow: In the estimation of mean for data from Lognormal distribution (Figure 2.21), using BC_a confidence intervals slightly worsens the Banks-B and Hutson-B performance in estimation, NPI-B is the least affected bootstrap method and Efron-B is not notably better performing, it is still the worst performing bootstrap method in the estimation of mean for small samples. In the estimation of variance for data from Lognormal distribution (Figure 2.22), using BC_a confidence intervals worsens the performance of NPI-B, Banks-B and Hutson-B (from the perspective of both metrics of assessment) and it does not improve the Efron-B performance in the estimation of variance for small samples.

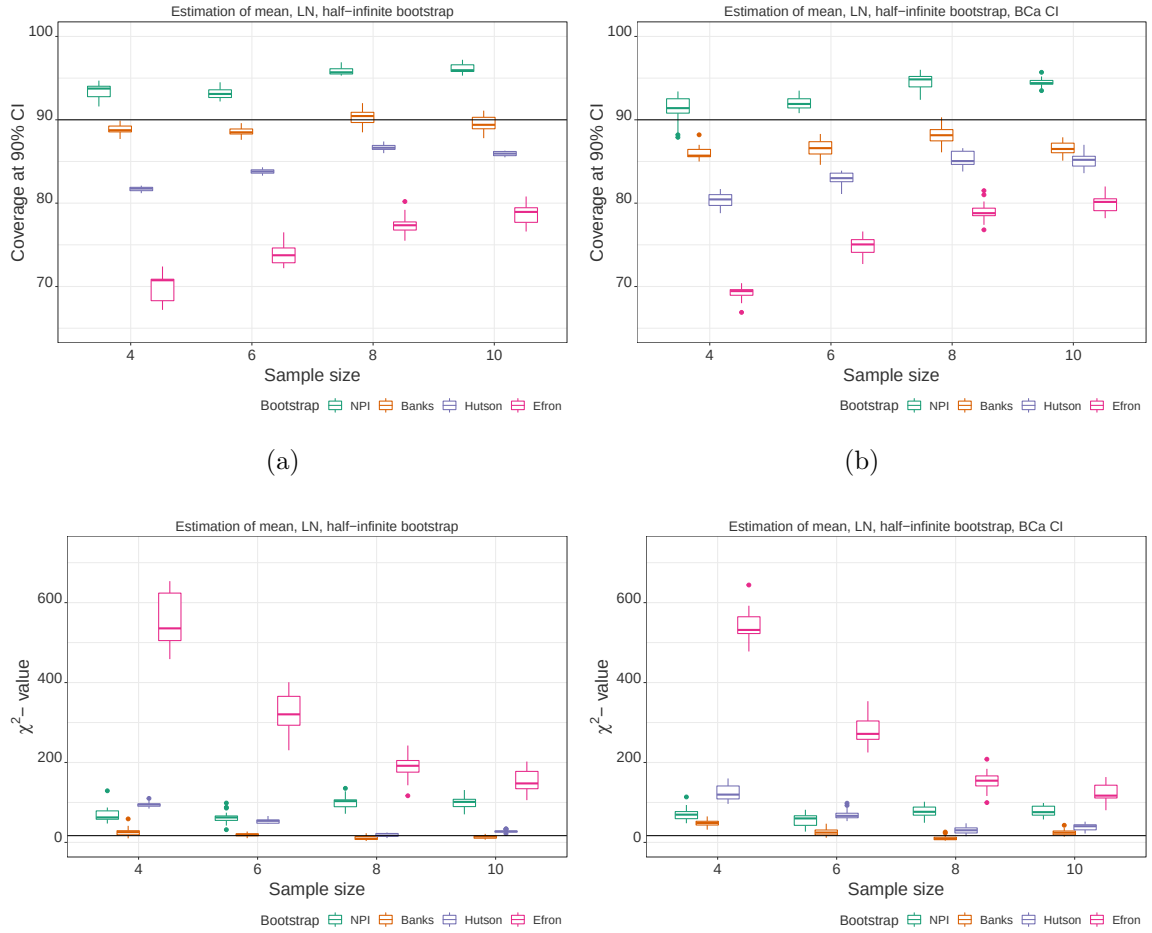


Figure 2.21: Coverage at 90% CI and χ^2-values, estimation of mean, LN($m_{LN} = -0.347, s^2_{LN} = 0.833^2$), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, percentile versus BC$_a$ confidence intervals, 20 simulations

Efron-B with BC$_a$ confidence intervals does not perform better in the estimation of quantiles (Q1, median, Q3) and NPI-B is not notably affected by the change of confidence intervals (see Figures 2.23, 2.24 and 2.25). The performance of Banks-B and Hutson-B in the estimation of quantiles is affected by the choice of confidence intervals, in some cases using BC$_a$ confidence intervals improves the performance of these bootstrap methods, in other cases it worsens it. In the estimation of Q1 for data from Lognormal distribution (Figure 2.23), Banks-B and Hutson-B have better coverage at 90% CI and lower χ^2-value with BC$_a$ confidence intervals for $n = 4, 8$. By contrast, for $n = 6, 10$ these two bootstrap methods have slightly higher χ^2-values. In the estimation of median for data

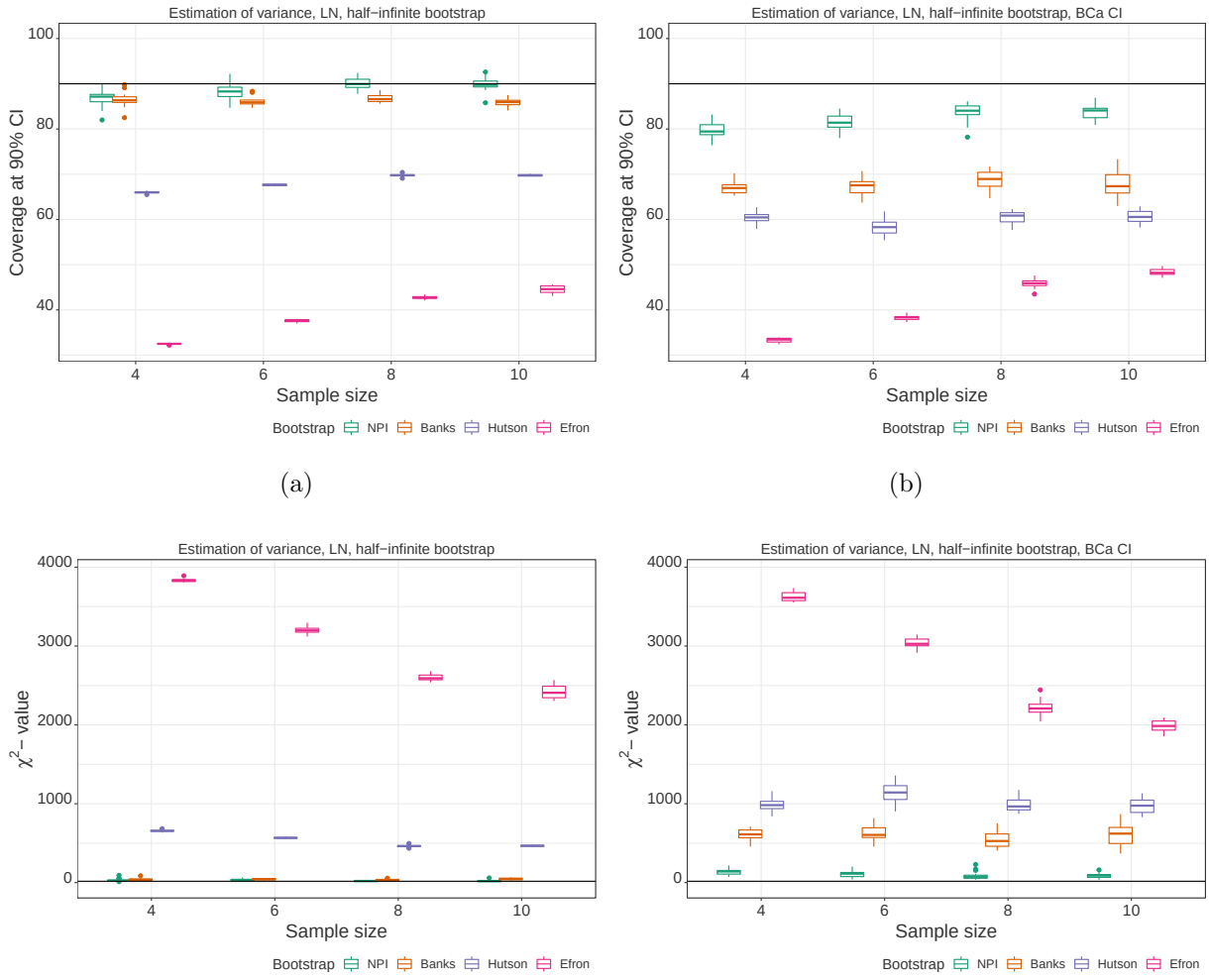


Figure 2.22: Coverage at 90% CI and χ^2-values, estimation of variance, LN($m_{LN} = -0.347, s^2_{LN} = 0.833^2$), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

from Lognormal distribution (Figure 2.24), using BC_a confidence intervals worsens the Banks-B's performance for $n = 4$ and the Hutson-B's performance for $n = 4, 10$. Although there is lower χ^2-value for $n = 6$ for both bootstrap methods and for $n = 8$ for Banks-B, this thesis does not recommend the use of BC_a confidence intervals in the estimation of median for either of the four bootstrap methods. In the estimation of Q3 for data from Lognormal distribution (Figure 2.25), using BC_a confidence intervals improves χ^2-value for Banks-B for all n and for Hutson-B for $n = 6, 8$. However, using BC_a CI also worsens coverage at 90% confidence interval for $n = 4$ for these two bootstrap methods and it

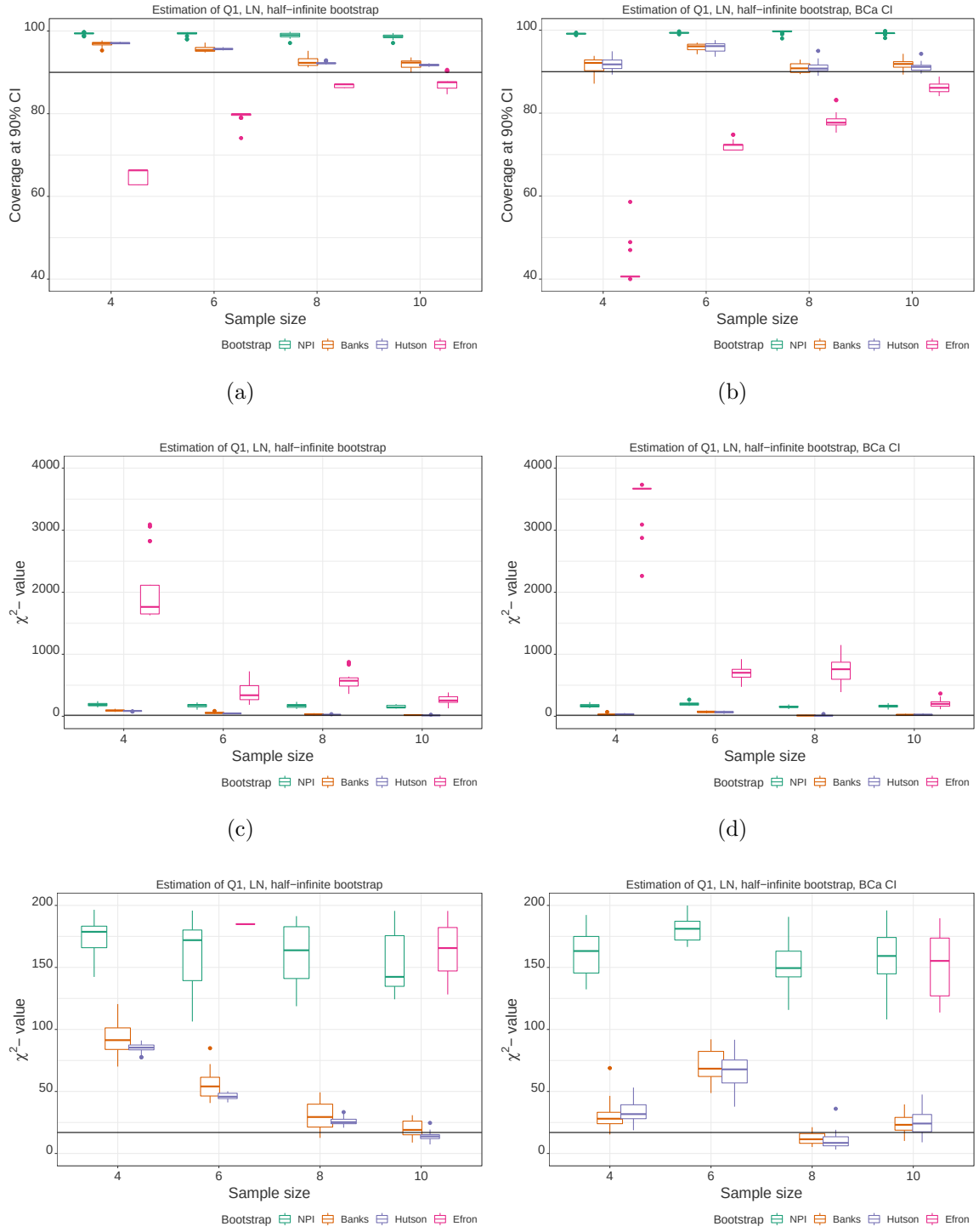
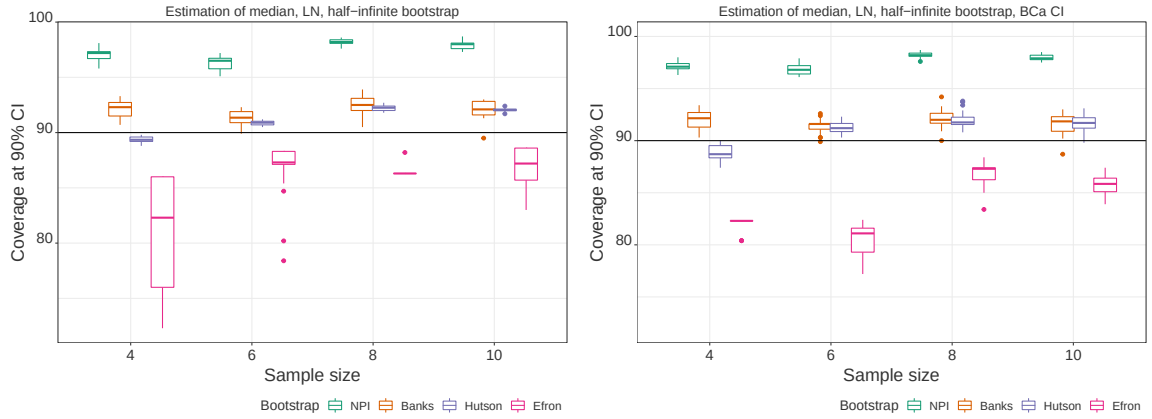
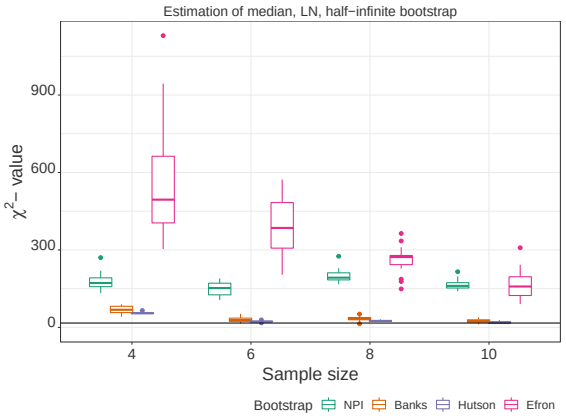


Figure 2.23: Coverage at 90% CI and χ^2-values, estimation of Q1, LN($m_{LN} = -0.347, s^2_{LN} = 0.833^2$), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

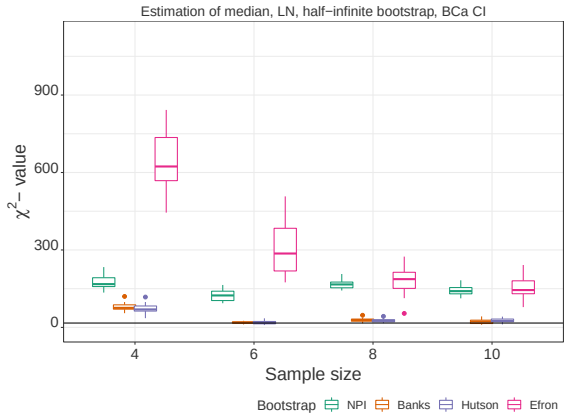


(a)

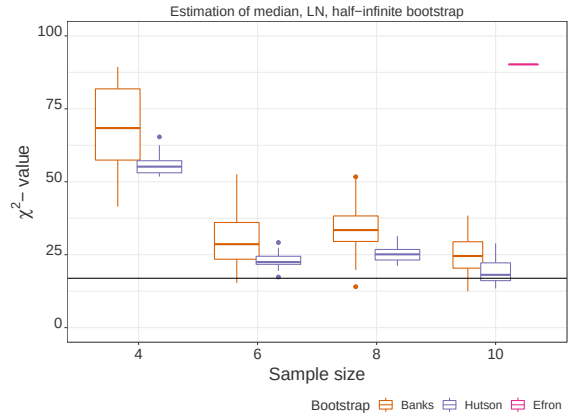
(b)



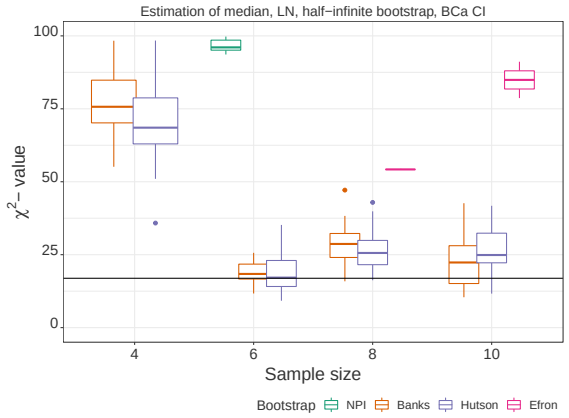
(c)



(d)



(e)



(f)

Figure 2.24: Coverage at 90% CI and χ^2-values, estimation of median, LN($m_{LN} = -0.347, s^2_{LN} = 0.833^2$), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, percentile versus BC$_a$ confidence intervals, 20 simulations

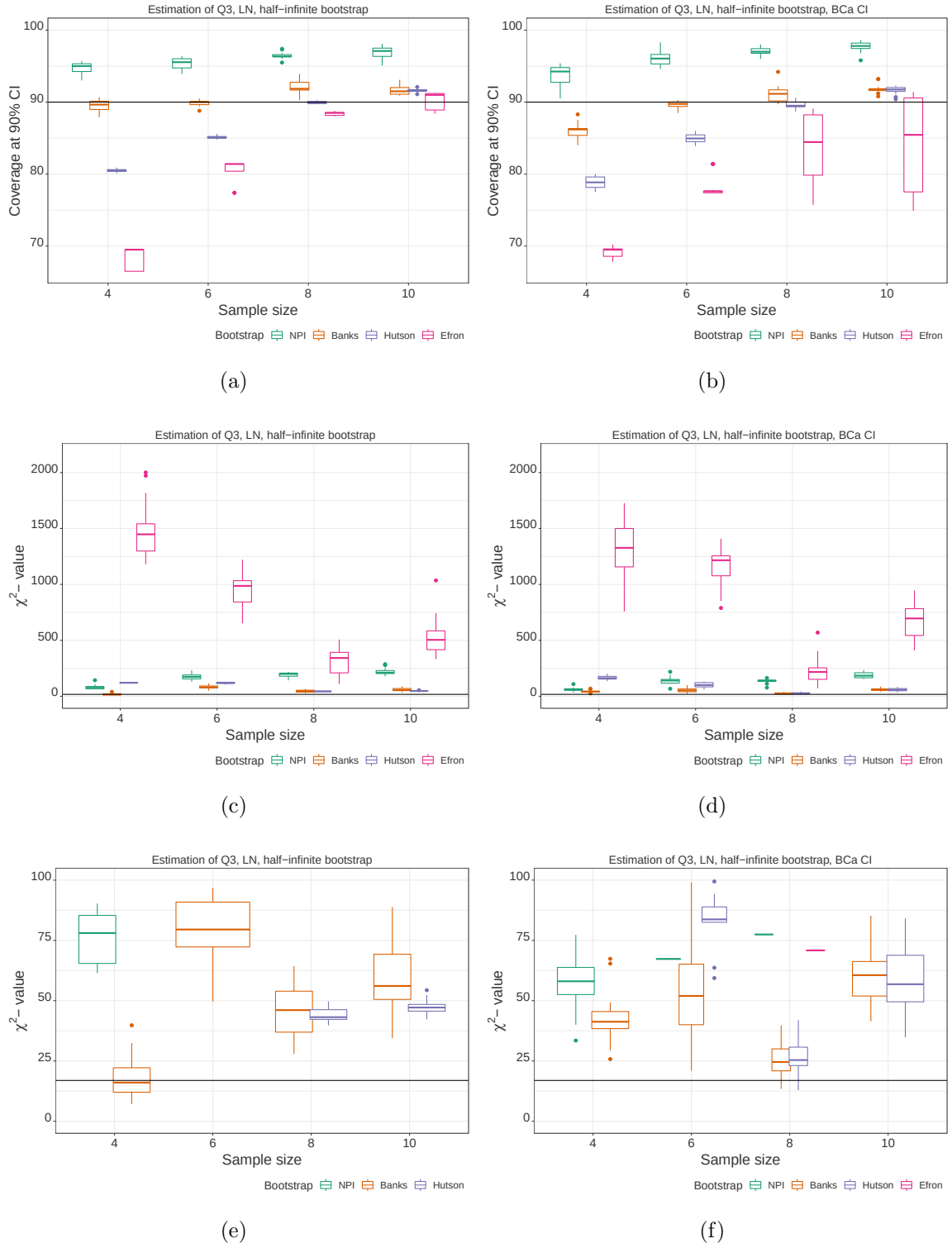


Figure 2.25: Coverage at 90% CI and χ^2-values, estimation of Q3, LN($m_{LN} = -0.347, s^2_{LN} = 0.833^2$), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

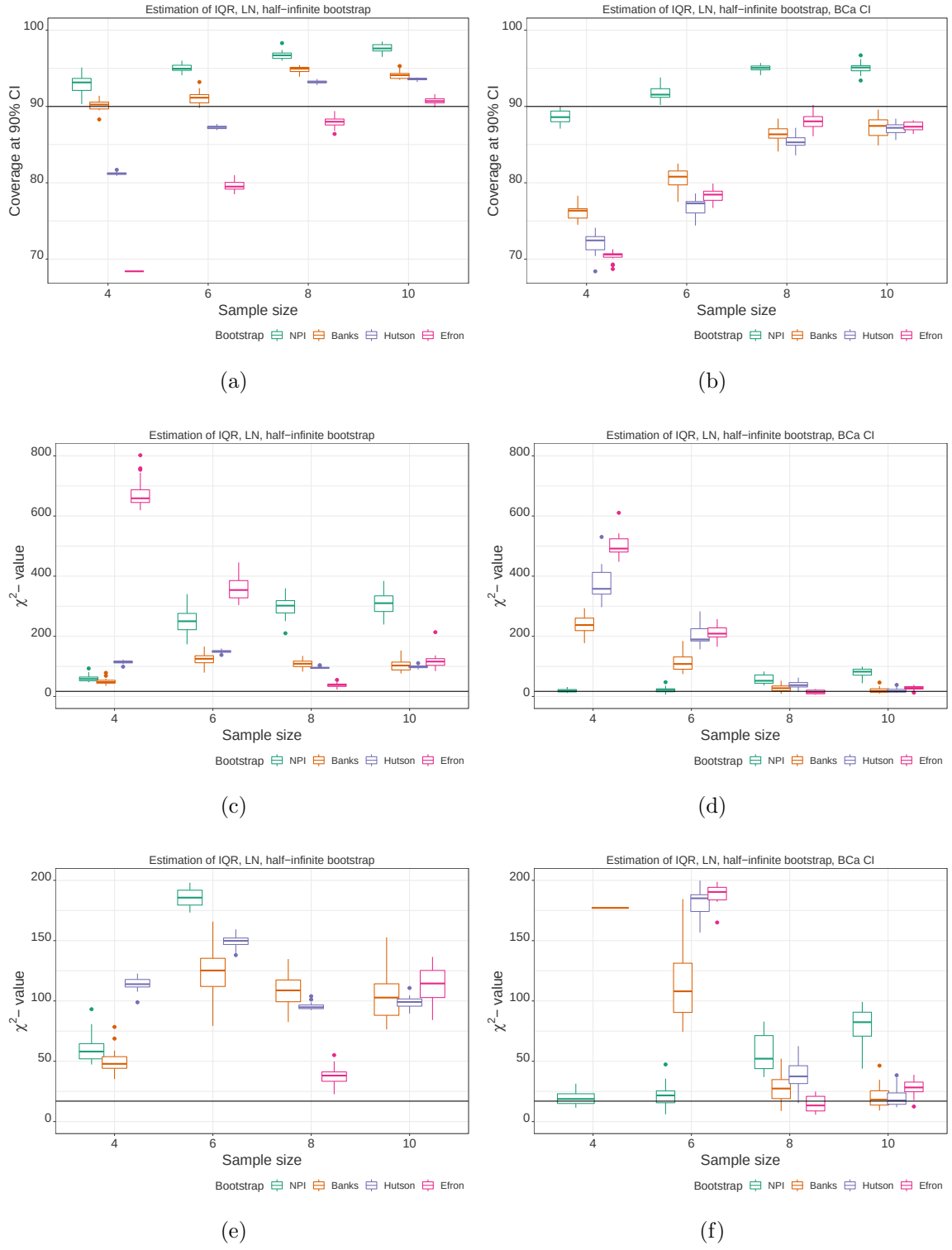


Figure 2.26: Coverage at 90% CI and χ^2-values, estimation of IQR, LN($m_{LN} = -0.347, s^2_{LN} = 0.833^2$), $n = 4, 6, 8, 10$, half-infinite NPI and Banks-B, percentile versus BC_a confidence intervals, 20 simulations

does not improve coverage for $n = 6, 8, 10$.

In the estimation of IQR for data from Lognormal distribution (Figure 2.26), NPI-B performs better with BC_a confidence intervals (lower χ^2-value for all n, better coverage at 90% confidence interval for $n = 6, 8, 10$, yet for $n = 4$, there is worse coverage – slight under-coverage). Banks-B and Hutson-B have worse coverage at 90% confidence interval for all n with BC_a confidence intervals. Although the χ^2-value is improved for $n = 6, 8, 10$ for Banks-B and for $n = 8, 10$ for Hutson-B, this work does not recommend using BC_a confidence intervals for Banks-B and Hutson-B in the estimation of IQR for Lognormally distributed data. For Efron-B, coverage at 90% confidence interval is not improved by using BC_a confidence intervals but χ^2-value is lower for all n. Thus, the pattern in the estimation of IQR is changed when BC_a confidence intervals are used instead of percentile confidence intervals, however, there are still no clear conclusions about the bootstrap method performance in the estimation of IQR.

Overall, the findings of this thesis dictate a recommendation of not using BC_a confidence intervals in the estimation of mean, variance and median for Lognormally distributed data for either of the four bootstrap methods. Further research into using BC_a confidence intervals with NPI-B and Efron-B in the estimation of IQR, and with Hutson-B and Banks-B in the estimation of Q1 and Q3 for Lognormally distributed data, is encouraged. Caution is advised to practitioners regarding the usage of BC_a confidence intervals in the estimation of quantiles and IQR before more research has been carried out.

2.4.4 Mixed-Normally distributed data

In practical applications, scientific community sometimes assumes Normal distribution in cases where sample size is low. But this is not always a fair assumption. The data can also form more than one Normal distribution. To account for this scenario, the performance in estimation is explored for data from Mixed-Normal distributions.

Determining the parameters of the Mixed-Normal distribution in a way that reflects the real life test scenarios is complicated. For example, in oncology clinical research, a meaningful response is 30% tumour reduction and at 20% tumour increase the patient is considered to be clinically not responding to the treatment. One option would be to use the clinical research circumstances as a starting point when creating Mixed-Normal

distributions. However, the problem is that in clinical research, humans already have a tumour, whereas in animal study there are healthy animals without tumour, which get inserted the tumour and are subsequently treated. In preclinical research when a group of animals is given a drug, some animals respond whilst others do not. In many cases this will be in the ratio 50 : 50. An example of this is the Mixed-Normal distribution A, $0.5N(0,1) + 0.5N(3,1)$. This Mixed-Normal distribution is formed from two Normal distributions with equal probability, the two distributions have different means but same standard deviations. Of course, this is only an attempt to reflect the real life scenario. Other Normal mixed ratio, 90:10, has been considered for two different Mixed-Normal distributions. In practical scenarios, Mixed-Normal distribution composed of two distributions with same means but different variances can occur when two different measuring tools with different inaccuracies are used. An example of this case is Mixed-Normal distribution B, $0.9N(0,1)+0.1N(0,16)$. In theory, Mixed-Normal distribution can also consist of two Normal distributions with different means and different variances. An example of this is Mixed-Normal distribution C, $0.9N(0,1)+0.1N(4,9)$. There is no particular reason for choosing those exact parameters for the Mixed-Normal distribution C, the important part is that the means and variances are sufficiently different. Plot of density functions for the Mixed-Normal distributions A, B and C is displayed in Figure 2.27.

$N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ are two independent Normal distributions. The distribution parameters for the Mixed-Normal distributions are: mean for $p_1N(\mu_1, \sigma_1^2) + p_2N(\mu_2, \sigma_2^2)$ is equal to $p_1\mu_1 + p_2\mu_2$ and variance for $p_1N(\mu_1, \sigma_1^2) + p_2N(\mu_2, \sigma_2^2)$ is equal to $p_1\sigma_1^2 + p_2\sigma_2^2 + p_1\mu_1^2 + p_2\mu_2^2 - \mu^2$, where μ is the mean of $p_1N(\mu_1, \sigma_1^2) + p_2N(\mu_2, \sigma_2^2)$. Thus, true mean for Mixed-Normal A $(0.5N(0,1) + 0.5N(3,1))$ is 1.5 and true variance for Mixed-Normal A is 3.25; true mean for Mixed-Normal B (0.9N(0,1)+0.1N(0,16)) is 0 and true variance for Mixed-Normal B is 2.5; true mean for Mixed-Normal C (0.9N(0,1)+0.1N(4,9)) is 0.4 and true variance for Mixed-Normal C is 3.24.

Clearly, there are limitations to estimation analysis for mixtures of Normal distributions when the sample sizes are small, especially when the sample size $n = 4$, because of the known identifiability issues with mixtures of Normal distributions. Unless, there is some biological or other knowledge that would suggest that the data could come from a Mixed-Normal distribution, the practitioner might not realise that a very small sample

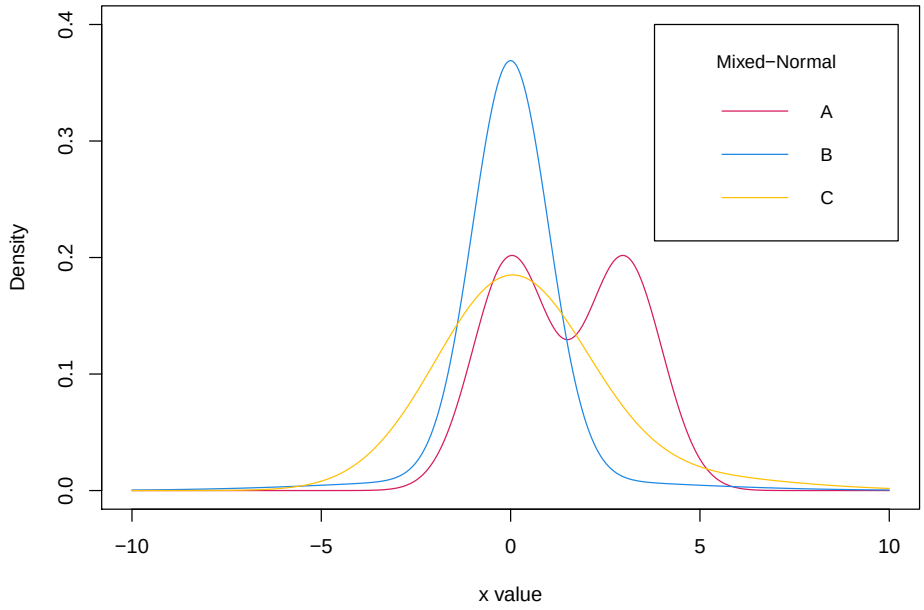


Figure 2.27: Plot of density functions for the Mixed-Normal distributions A, B and C

comes from such distribution. However, the point of the simulation study presented in this section is to observe how in such cases the bootstrap methods perform when estimating a particular population characteristics.

In the simulation study, finite range is assumed for Banks-B and NPI-B, and Hutson-B is applied on $(-\infty, \infty)$. In Appendix A.4.3, the choice of finite versus infinite range is briefly studied. Plots for Mixed-Normal distributions A, B and C for the estimation of mean and variance are displayed in Figures 2.28 and 2.29, respectively.

For the Mixed-Normal distribution A, it can be concluded that for both the estimation of mean and variance, Banks-B is a good choice of the bootstrap method as it has low χ^2-value and from n=6 onwards it also has small over-coverage at 90% CI. Banks-B has small under-coverage at $n = 4$. Hutson-B has better coverage at 90% CI and lower χ^2-values in the estimation of mean for $n = 4$ and better coverage at 90% CI in the estimation of variance for $n = 4$.

For the Mixed-Normal distribution B, it can be concluded that Banks-B performs well in the estimation of mean as there is good coverage at 90% CI and the lowest χ^2-value but not so well in the estimation of variance. For the estimation of variance for the

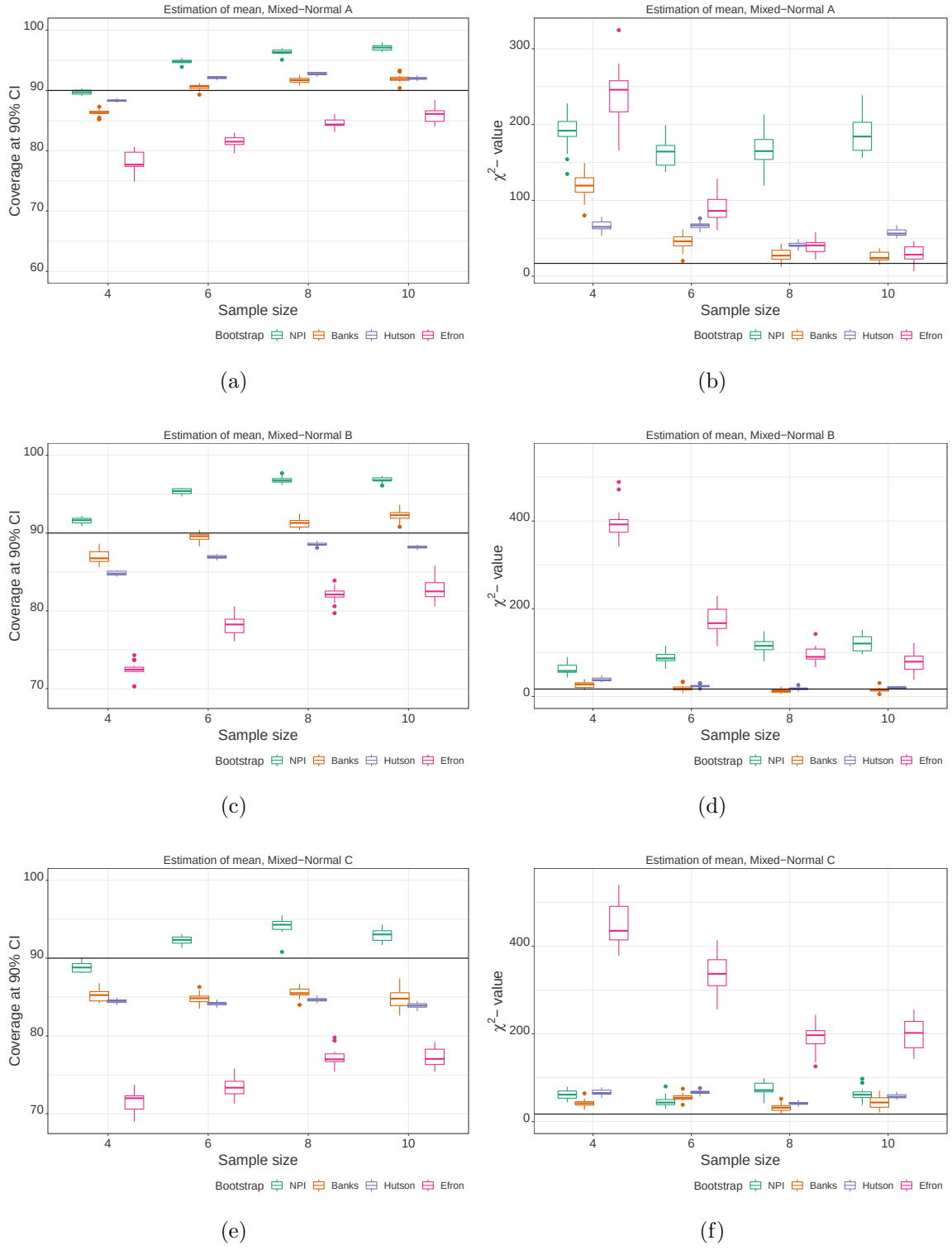


Figure 2.28: Coverage at 90% CI and χ^2-values, estimation of mean, Mixed-Normal A, B, C, $n = 4, 6, 8, 10$, finite (Approach I) Banks-B and NPI-B, 20 simulations

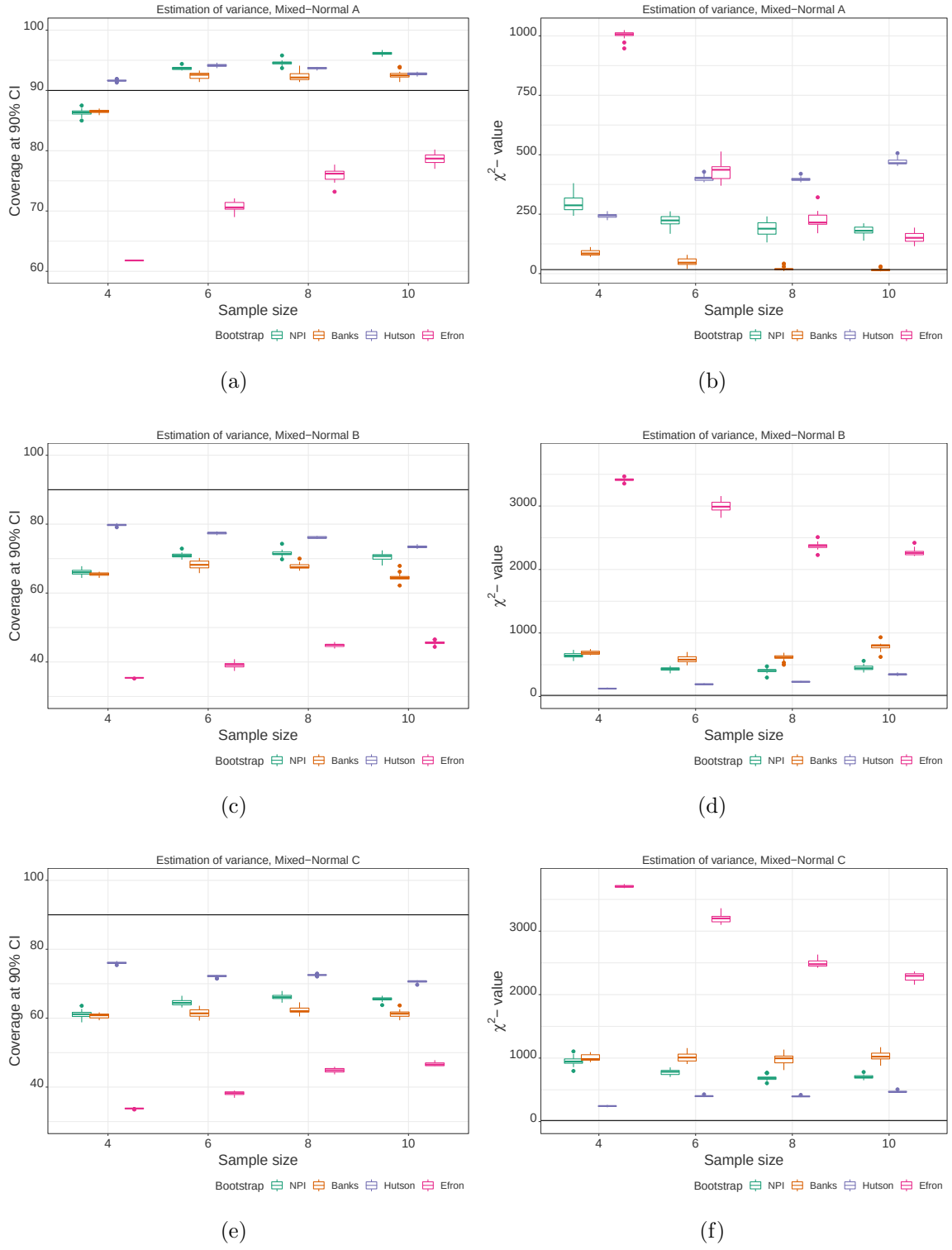


Figure 2.29: Coverage at 90% CI and χ^2-values, estimation of variance, Mixed-Normal A, B, C, $n = 4, 6, 8, 10$, finite (Approach I) Banks-B and NPI-B, 20 simulations

Mixed-Normal distribution B, all the bootstrap methods have under-coverage, Efron-B has the largest under-coverage and Hutson-B has the smallest under-coverage. Hutson-B performs the best, and NPI-B the second best, out of the four bootstrap methods in the estimation of variance, but it is arguable whether this bootstrap method should be recommended as a tool for outlier situations, as there is still large under-coverage.

For the Mixed-Normal distribution C, it can be concluded that NPI-B has the best coverage in the estimation of mean and on average little bit larger χ^2-value than Banks-B. Thus, it is arguable whether NPI-B or Banks-B perform better. Hutson-B is the third best performing bootstrap method in the estimation of mean for Mixed-Normal distribution C. In the estimation of variance for Mixed-Normal distribution C, Hutson-B is the best performing bootstrap method, as it has the lowest χ^2-value, but it still has under-coverage at 90% CI. Therefore, the decision-maker should still be careful about making choices based on estimation in this case. Efron-B performs the worst for both the estimation of mean, and variance for all three Mixed-Normal distributions.

Overall, Banks-B is the best bootstrap method choice for the estimation of mean and variance for Mixed-Normal distribution consisting of two distributions of different means, but same variance; Hutson-B performs the best in the estimation of variance for two Mixed-Normal distributions consisting of two different distributions with different variances; Efron-B is the worst choice for the estimation of mean and variance for data from Mixed-Normal distribution for small samples. However, making conclusions and decisions based on these estimations should be done with care. Appendix A.4.3 briefly compares the use of infinite range and finite range and concludes that infinite range is preferable.

2.4.5 Summary

This section summarises findings of the bootstrap methods' performance in the quantification of uncertainty for small samples in the estimation of mean, variance and quantiles for Normally, Lognormally, Exponentially and Mixed-Normally distributed data. The simulation study has shown that Efron-B performs poorly in the estimation of mean, variance and quantiles for small samples. Across different distributions, there is large under-coverage at 90% confidence interval and large χ^2-value when estimating mean,

variance and quantiles for small sample sizes ($n \leq 10$). This conclusion is in line with the reasoning of the asymptotic argument introduced in Section 2.1.

Moreover, the study has shown that Banks-B performs better than Efron-B in the quantification of uncertainty for small samples in the estimation of mean, variance and quantiles for Normally, Lognormally, Exponentially and Mixed-Normally distributed data, when sample sizes are small. Even using BC_a confidence intervals did not make Efron-B better performing than Banks-B in estimation for small samples for both Normally and Lognormally distributed data.

Hutson-B performs well in the estimation of variance of Normally distributed data and Mixed-Normally distributed data consisting of two distributions with different variances, but not so well in the estimation of variance of Lognormally and Exponentially distributed data. Even though Hutson-B performs well in the estimation of variance for Normally distributed data, this thesis would not recommend it for the estimation of variance of small samples because in such cases the underlying distribution cannot be ascertained. Hutson-B also performs well in the estimation of quantiles for Normally distributed data and in some cases even for the estimation of quantiles for Lognormally and Exponentially distributed data.

NPI-B is aimed at prediction, not estimation, nevertheless, this investigation has shown that there are cases where NPI-B performs well in estimation. Generally, NPI-B performs the best in the estimation of population characteristics for $n = 4$ and in the estimation of variance for skewed data.

Table 2.1 summarises recommendations for the bootstrap method choice when estimating various population characteristics for small sample size ($n \leq 10$) for different choices of distributions. When estimating population characteristics for small samples, researchers often have limited knowledge about the underlying distribution of the data. On balance, despite this limited knowledge, the investigation has shown that in most cases, across different cases, Banks-B is a good choice of the bootstrap method for the estimation of population characteristics for data with small sample size ($n \leq 10$).

Distribution	Metric	Best performing bootstrap
Normal	mean	finite Banks-B or Hutson-B
Normal	variance	Hutson-B
Normal	quantiles	finite Banks-B or Hutson-B
Normal	IQR	no clear conclusion
Lognormal or Exponential	mean	half-infinite Banks-B
Exponential	variance	half-infinite Banks-B
Lognormal	variance	half-infinite NPI-B
Lognormal or Exponential	quantiles	half-infinite Banks-B
Lognormal or Exponential	IQR	no clear conclusion
Mixed-Normal different means, same variances	mean	infinite Banks-B
Mixed-Normal same means, different variances	mean	infinite Banks-B
Mixed-Normal different means, different variances	mean	infinite Banks-B or NPI-B
Mixed-Normal different means, same variances	variance	infinite Banks-B
Mixed-Normal same means, different variance	variance	Hutson-B
Mixed-Normal different means, different variances	variance	Hutson-B

Table 2.1: Summary table of the comparison cases considered for the bootstrap performance in estimation for small sample size ($n \leq 10$)

The simulation study for Normally distributed data showed that both finite and infinite range may be employed for Banks-B and NPI-B in the estimation of mean and quantiles, however finite bootstrap is computationally easier. In the estimation of variance for Normally distributed data, the infinite bootstrap is the recommended range. When the data come from Exponential or Lognormal distribution, the choice of half-infinite Banks-B is recommended for the estimation of mean and quantiles (Q1, median and Q3) when the sample size is $n = 10$ or smaller. This work would recommend half-infinite NPI-B for the estimation of variance for Lognormally distributed data, or data from another skewed distribution, and half-infinite Banks-B for the estimation of variance for data with underlying Exponential distribution.

This investigation did not provide explicit conclusions about the bootstrap performance in the estimation of IQR. The performance of the bootstrap methods in the estimation of IQR differed per sample size and per distribution. Thus, this remains an open topic for future research.

Furthermore, this study explored the effect of using BC_a confidence intervals instead of percentile confidence interval for Normally and Lognormally distributed data. BC_a CI did not make Efron-B a better performing bootstrap method than the other three bootstrap methods (Banks-B, Hutson-B and NPI-B) from the perspective of either of the two metrics of assessment for both Normally and Lognormally distributed data. On the other hand, for Normally distributed data of small sample size, further research into BC_a confidence for Hutson-B for the estimation of mean and quantiles (Q1, median and Q3) and for Banks-B and NPI-B for the estimation of quantiles, is recommended. This work would not recommend BC_a confidence intervals for the estimation of mean, variance and median for Lognormally distributed data for either of the four bootstrap methods. But further research into using BC_a confidence intervals with NPI-B and Efron-B for the estimation of IQR, and with Hutson-B and Banks-B for the estimation of Q1 and Q3 for Lognormally distributed data, is encouraged. Caution is advised to practitioners about using BC_a confidence intervals instead of percentile confidence for the estimation of any population characteristics for small samples, before more research has been carried out.

Some further work regarding the bootstrap method performance in estimation for small samples has not been included in the main investigation, but it is reported in

Appendix. The bootstrap method was briefly studied for smaller sample sizes, $n = 2, 3$, for Normally distributed data and the findings showed that Banks-B, Hutson-B and NPI-B performed much better than Efron-B for these sample sizes. However, this thesis would not recommend the bootstrap method for such small sample sizes. The initial remarks for Normally distributed data can be found in Appendix A.2.

Moreover, a brief investigation into smoothed bootstrap using kernel (Kernel-B), presented in Appendix A.5.2, showed that Kernel-B performs decently well in the estimation of quantiles (Q1, median and Q3). However, further research into the use of smoothing parameter is recommended before employing Kernel-B in practice for small samples.

2.5 Bootstrap coverage performance in prediction

This section assesses the performance in making prediction inference for small sample sizes for four bootstrap methods (NPI-B, Banks-B, Hutson-B and Efron-B) for data simulated from Normal, Lognormal, Exponential and Mixed-Normal distributions. The prediction coverage at 90% percentile prediction intervals, when estimating mean, variance, median, Q1, Q3 and IQR of small sample sizes ($n = 4, 6, 8, 10, 20$), is considered. Initial investigation of the bootstrap methods' performance in prediction can be found in Coolen and BinHimd [53] and in BinHimd's thesis [31]. BinHimd compared the performance in the prediction of mean, variance and Q3 for NPI-B and Efron-B, focusing on larger samples ($n = 20, 50, 100, 200, 500$). She considered 0.90% and 0.95% prediction interval and Normal, Gamma and Uniform distribution. This study extends the previous exploration, focusing on small samples and including two additional bootstrap methods, Banks-B and Hutson-B.

The simulation study used to assess the bootstrap performance in prediction is outlined in Section 2.5.1 and the findings about the bootstrap methods' performance in making prediction inference are presented for Normal, Lognormal, Exponential and Mixed-Normal distributions in Sections 2.5.2, 2.5.3 and 2.5.4, respectively.

Algorithm 2 Calculating the prediction coverage of the $100(1-2\alpha)\%$ percentile prediction interval

1: Draw $2N$ samples of size n from a specific distribution. Consider $X_1, X_2, \ldots, X_N$ as the actual samples and $X_{N+1}, \ldots, X_{2N}$ as the future samples.
2: For each future sample calculate a chosen sample statistic $\hat{\theta}_i$, $i \in \{1, \ldots, N\}$.
3: From each actual sample i, create B bootstrap samples of size n, construct $100(1-2\alpha)\%$ prediction interval for $\hat{\theta}_i$, and record whether $\hat{\theta}_i$ is in the percentile prediction interval.
4: Report the proportion (in %) of intervals which contain $\hat{\theta}_i$ out of the N prediction intervals.
5: In total carry out Steps 1-4 M times.

2.5.1 Methodology

The simulation for assessing the bootstrap performance in prediction is described in Algorithm 2. The algorithm for assessing the prediction performance of bootstrap methods is based on BinHimd's algorithm presented in [31]. Let $X = (X_1, X_2, \ldots, X_n)$ be the actual random samples and $X^* = (X_{n+1}, \ldots, X_{2n})$ be the future random samples, both X and X^* are iid from the same probability distribution. Percentile prediction intervals are used to predict the statistic $\hat{\theta}$, e.g. mean. From each actual sample i, $i \in \{1, \ldots, N\}$, B bootstrap samples are created: $y^{*i1}, y^{*i2}, \ldots, y^{*iB}$. For each bootstrap sample b, $b \in \{1, \ldots, B\}$, the bootstrapped sample statistic, $\hat{\theta}^*_{ib}$, is calculated. The $100(1-2\alpha)\%$ prediction interval for the population parameters is constructed by defining the lower bound to be the αB^{th} value in the ordered list of the $\hat{\theta}^*_{ib}$ values and the upper bound to be the $(1-\alpha)B^{th}$ value in this list. For each future sample, $\hat{\theta}_i$ is calculated and it is assessed whether this value lies in the percentile prediction interval of the actual sample.

In this investigation, $B = 1000$, $N = 1000$, $M = 20$, $\alpha = 0.05$. The study focuses on the coverage at 90% percentile prediction interval. In the figures, PPI stands for percentile prediction interval. The best performing bootstrap has coverage of 90% at 90% PPI. Over-coverage is better than under-coverage and the explanation for that is similar to the one for the bootstrap performance in estimation outlined in Section 2.4.1.

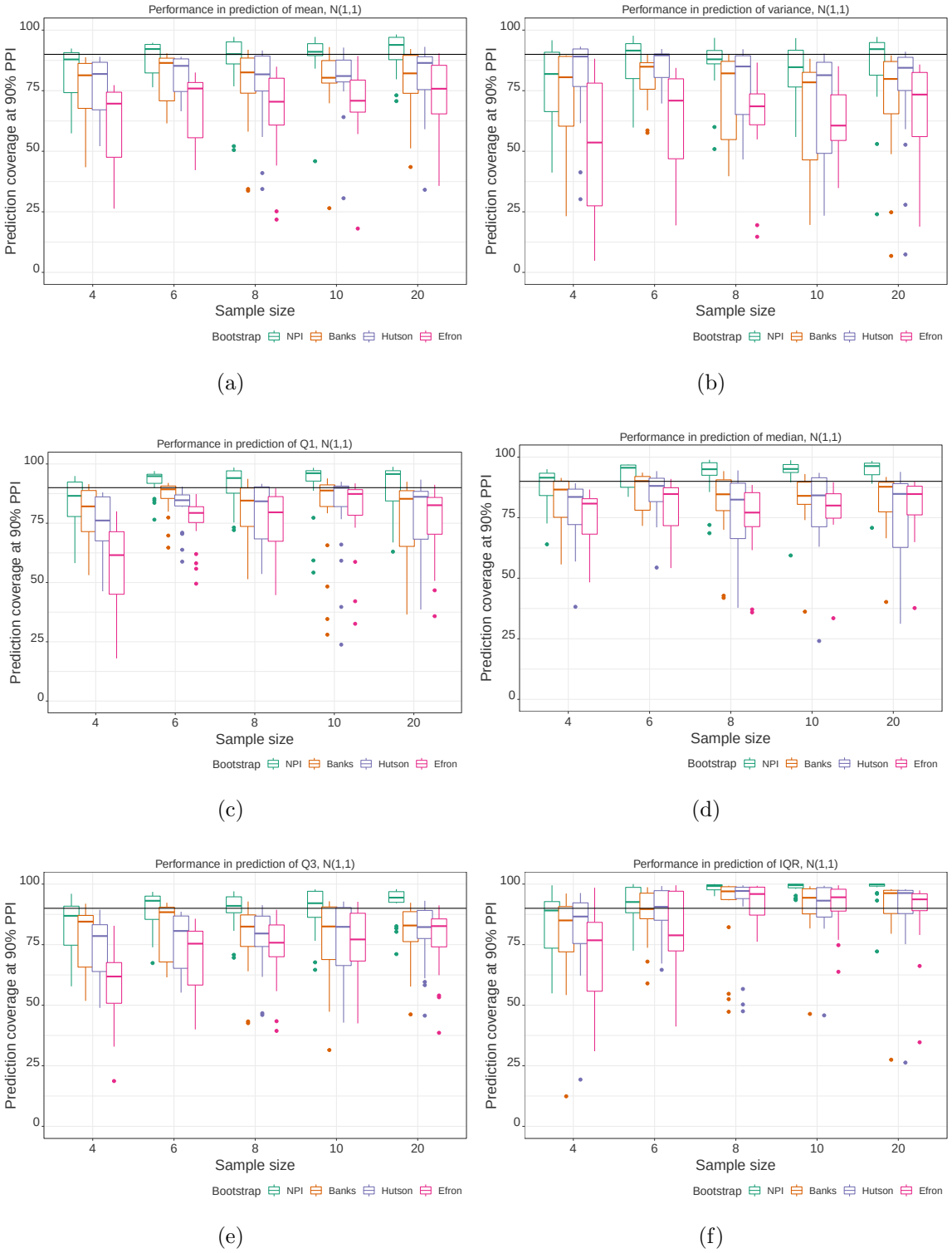


Figure 2.30: Prediction performance of bootstrap methods for data from N(1,1), for prediction of mean, variance, quantiles and IQR, finite NPI-B and Banks-B, 20 simulations

2.5.2 Normally distributed data

The prediction performances for mean, median, Q1, Q3, variance and IQR for Normally distributed data are visualised in Figure 2.30. Finite range for NPI-B and Banks-B (Approach I, Section 2.3.3) is used. From the simulations it can be concluded that for the performance of mean, variance and the quantiles (Q1, Q2 and Q3), NPI-B performs the best in prediction, as for other bootstrap methods, there is under-coverage, and for NPI-B there is mostly over-coverage and good coverage in some cases. Hutson-B performs better in the prediction of variance at $n = 4$. This could be linked to the large variance of Hutson-B variance outputs (see Figure A.8, Appendix A.3).

In the prediction of IQR, all bootstrap methods perform well for $n = 8, 10, 20$. For $n = 4$, NPI-B performs the best in the prediction of IQR and at $n = 6$, NPI-B, Banks-B and Hutson-B perform well. The good performance in the prediction of IQR of all bootstrap methods is a consequence of the Normal distribution. NPI-B assumes more variability, as the data are usually not perfectly Normally distributed. The large over-coverage of NPI-B can be a result of NPI accounting for larger variability of data.

2.5.3 Exponential and Lognormal distributions

The prediction performances for mean, median, Q1, Q3, variance and IQR for Exponentially and Lognormally distributed data are displayed in Figures 2.31 and 2.32, respectively. Half-infinite range (Approach V, Section 2.3.3), has been used for NPI-B and Banks-B for Exponentially and Lognormally distributed data. The simulation study shows that NPI-B performs well in prediction for Exponentially and Lognormally distributed data. Moreover, the prediction coverage of NPI-B for Exponentially and Lognormally distributed data is better than for Normally distributed data. This could be explained by the fact that NPI-B has bigger variance than other bootstrap methods, because it aims to embrace uncertainty and there is more uncertainty (unpredictability) in Exponentially and Lognormally distributed datasets than in Normally distributed ones. There are some cases where Banks-B or Hutson-B perform well. Namely, for $n = 4$ in the prediction of variance for Exp(1), Banks-B has better coverage than NPI-B, and for $n = 4, 6$ for the prediction of Q1 for Exp(1), Banks-B and Hutson-B have the best coverage. For

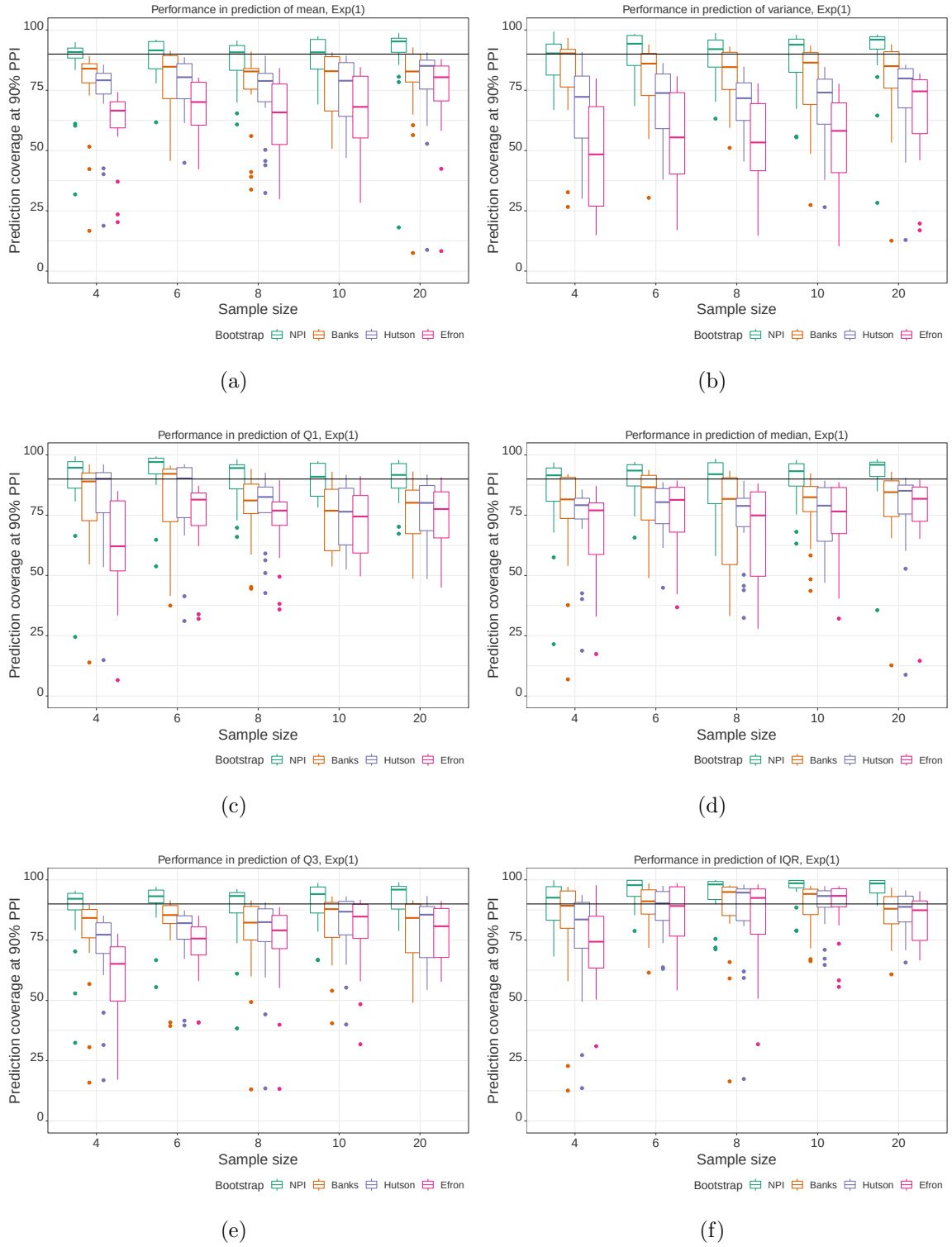


Figure 2.31: Bootstrap methods performance in prediction of mean, variance, quantiles and IQR for data from Exp(1), 20 simulations

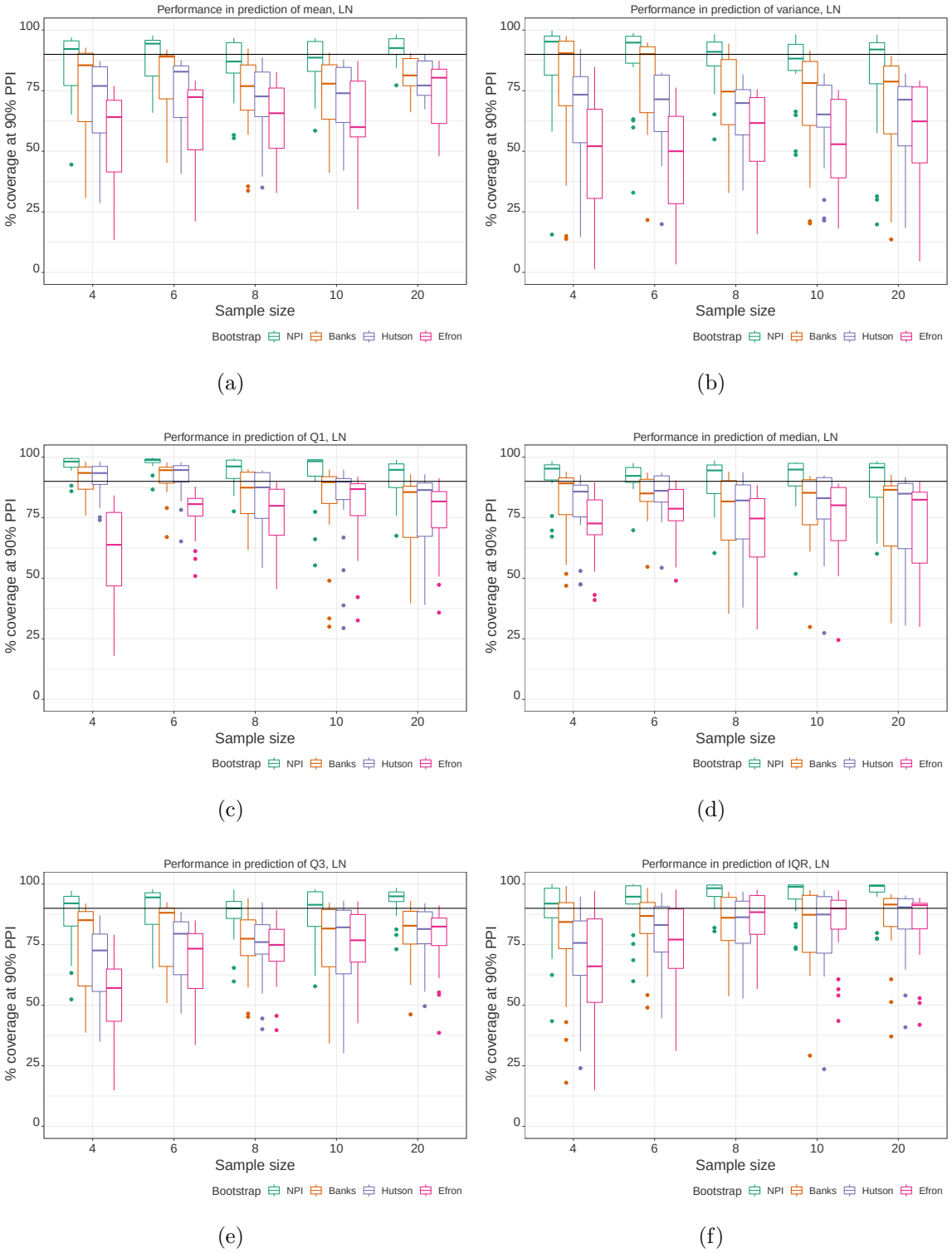


Figure 2.32: Bootstrap methods performance in prediction of mean, variance, quantiles and IQR for data from Lognormal ($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$), 20 simulations

Lognormally distributed data, Banks-B and Hutson-B perform well in the prediction of Q1 for $n = 4, 6$ and Banks-B in the prediction of median for $n = 4, 6$. Efron-B is the worst performing bootstrap method in the prediction of mean, variance and quantiles. Banks-B performs well in the prediction of IQR for Exponentially distributed data for all studied sample sizes, Hutson-B and Efron-B for $n = 6, 8, 10$. In the prediction of IQR for Lognormally distributed data, Efron-B performs well for $n = 10, 20$, and Banks-B and Hutson-B for $n = 20$. Overall, this thesis recommends the use of NPI-B for prediction, especially in the case where it is not clear what the underlying distribution is.

2.5.4 Mixed-Normally distributed data

The prediction performance has also been explored for the four bootstrap methods for data generated from Mixed-Normal distribution A $(0.5N(0,1) + 0.5N(3,1))$, B $(0.9N(0,1) + 0.1N(0,16))$ and C $(0.9N(0,1) + 0.1N(4,9))$ for the prediction of mean and variance and the outputs of the simulation are displayed in Figure 2.33. The conclusion of this simulation is that NPI-B is the best performing bootstrap method and Efron-B is the worst performing bootstrap method in the prediction of both mean and variance for all three Mixed-Normal distributions, given that NPI-B has the best coverage and Efron-B has the largest under-coverage. NPI-B performs the best for Mixed-Normal distribution A, which consists of two Normal distributions with different means, but same variance. NPI-B is still the best performing bootstrap method in the prediction of mean and variance for Mixed-Normal distributions B and C, which consist of two different Normal distributions of different variances. It is possible that it is harder for NPI-B (and other bootstrap methods) to capture this aspect of the Mixed-Normally distributed data.

There are cases where even NPI-B has under-coverage. This is the case for only small samples for Mixed-Normal A ($n = 4, 6$ for the prediction of mean and $n = 4$ for prediction of variance). However, for the estimation of variance for Mixed-Normal distribution B and C, there is under-coverage for all sample sizes. In the estimation of mean for these two mixed distributions, NPI-B has good coverage only for sample sizes $n = 8, 10$, for the rest of the sample sizes, NPI-B has under-coverage. Nevertheless, NPI-B has coverage closest to 90% at 90% PPI, thus, it still performs the best.

Banks-B and Hutson-B perform better than Efron-B but worse than NPI-B in the

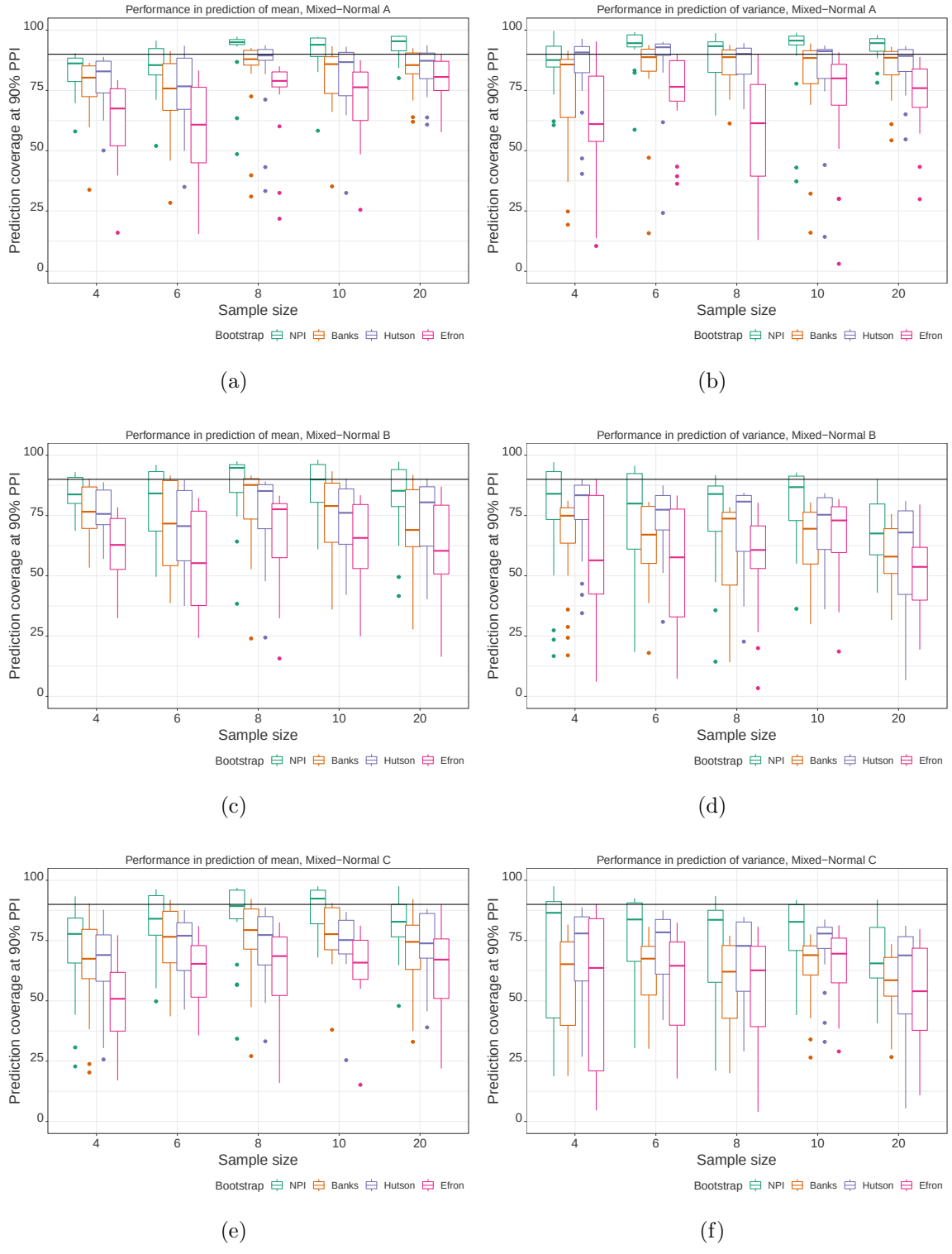


Figure 2.33: bootstrap methods performance in prediction of mean and variance for data from Mixed-Normal A, B and C, finite (Approach I) NPI-B and Banks-B, 20 simulations

prediction of mean and variance of Mixed-Normal data. Hutson-B is slightly better than Banks-B in the prediction of variance for all three Mixed-Normally distributed data and for the prediction of mean for Mixed-Normal distribution A. Banks-B and Hutson-B perform similarly for the prediction of mean for Mixed-Normal C and Banks-B performs better than Hutson-B in the prediction of mean for Mixed-Normal B. For more observations, further simulations are needed. It would be beneficial to explore more parameters for Mixed-Normal distribution.

2.5.5 Summary

The study concludes that NPI-B is superior to Banks-B, Hutson-B and Efron-B regarding the performance in the prediction of mean, variance and quantiles (Q1, median and Q3) for both Normally, Exponentially, and Mixed-Normally distributed data of small sample sizes ($n = 4, 6, 8, 10, 20$). However, NPI-B is not the best performing bootstrap in the prediction of IQR, especially for Normally distributed data for sample sizes $n = 8, 10, 20$, where Banks-B and Hutson-B perform better. A possible explanation for that is that their outputs are less varied than the outputs of NPI-B but more varied than outputs for Efron-B. Overall, Banks-B and Hutson-B perform better in prediction than Efron-B. Further research could explore the bootstrap method performance for data from a larger variety of distributions.

2.6 Bootstrap hypothesis testing

Sections 2.4 and 2.5 compared the bootstrap methods' performance in the estimation and prediction of population characteristics for small sample sizes. Section 2.2 introduced the use of Efron-B in bootstrap hypothesis testing for small samples. This section extends the exploration of bootstrap hypothesis testing via a simulation study that also employs Banks-B and NPI-B. Even though NPI-B is not meant for estimation, it is included in this study to investigate how the hypothesis testing outcomes would differ for this bootstrap method for small sample sizes, compared to Efron-B and Banks-B. The main question of interest is whether Banks-B performs well in bootstrap hypothesis testing. Hutson-B is not included in this example of an application because Banks-B is better in the estimation

of mean, especially for data that do not follow Normal distribution, as shown in Section 2.4.3, and in this example of hypothesis testing scenario the focus is on means.

Both bootstrap hypothesis test and permutation test can be used instead of conventional comparison tests. Both methods rely on the randomisations of the observed data [166]. The difference between the two methods is that bootstrap method quantifies the sampling distribution of some statistic computed from the data, whereas permutation test seeks to quantify the null distribution [166]. For a comparison of two samples, the limitation of a permutations test is that it requires that both groups both groups in the comparison have equal variance [210], whereas bootstrap methods do not have this requirement.

The algorithm for calculating the approximate p-value via the bootstrap hypothesis test, presented by Algorithm 3, is based on the algorithm presented by Dwivedi et al. [73]. This algorithm only works for two-sided comparison test, as explained in Section 2.2. In the algorithm, the unpooled variance is calculated, meaning that the Welsch t-test is used. The R code provided by the authors [73] is inconsistent with Algorithm 3 in one detail: in the R code, Student's t-test, which uses pooled variance, is used in Step 5 of Algorithm 3 instead of Welsch t-test (when comparing the two bootstrap samples). This thesis follows Algorithm 3 and the R code provided by Dwivedi et al. [73] was adjusted, in a way that unpooled variance is used for both the original and bootstrap samples. In Algorithm 3, B is set to $B = 1000$.

Original samples were generated from Normal, Lognormal and Skewed-Normal (using R package: `fGarch`, function: `rsnorm`) distributions. The study of Skewed-Normal distribution was limited to skewing parameter 0.8. Lognormal distributions employed were: $\text{LN}_A = \text{LN}(1,0.36)$, $\text{LN}_B = \text{LN}(2,1)$ and $\text{LN}_C = \text{LN}(3,16)$. The focus is on sample sizes $n = 5$ and $n = 10$ and on two original samples of unequal sample sizes ($n = 5$ vs. $n = 10$ and $n = 3$ vs. $n = 7$). Both finite (Approach I, Section 2.3.3) and infinite (Approach IV, Section 2.3.3) approaches are used for determining the first and the last interval for Banks-B and NPI-B. Type I error and estimated power are calculated for NPI-B bootstrap t-test (finite and infinite range), Banks-B bootstrap t-test (finite and infinite range), Efron-B bootstrap t-test, Student's t-test and Welsch t-test. The description of the methodology for calculating type I error and the estimated power will follow.

Algorithm 3 Calculating approximate p-value for bootstrap hypothesis test

1: Let $x = x_1, x_2 \ldots, x_{n_x}$ be the observations from sample X with mean $\overline{x}$ and sample standard deviation s_x and $y = y_1, y_2 \ldots, y_{n_y}$ be the observations from sample Y with mean $\overline{y}$ and sample standard deviation s_y;

2: Evaluate test statistic: $t_{\text{obs}} = \frac{\overline{x}-\overline{y}}{\sqrt{\frac{s_x^2}{n_x}+\frac{s_y^2}{n_y}}}$;

3: Combine samples x and y;

4: Draw two bootstrap samples from the combined sample: one of size n_x observations (x^*) and another of size n_y observations (y^*);

5: Compute mean and variance of each bootstrap sample as $(\overline{x}^*, \overline{s}_x^{*2})$ and $(\overline{y}^*, \overline{s}_y^{*2})$, respectively;

6: Evaluate test statistic: $t^* = \frac{\overline{x}^*-\overline{y}^*}{\sqrt{\frac{s_x^{2*}}{n_x}+\frac{s_y^{2*}}{n_y}}}$;

7: Repeat Steps 4-6 B times and obtain B values of the test statistic (t^*), i.e. $t^{*1}, t^{*2}, \ldots, t^{*B}$;

8: Approximate p-value= $\frac{1}{B}\sum_{j=1}^{B} \mathbb{1}_{(t^{*j} \geq t_{\text{obs}})}$

Type I error is the probability that the test will incorrectly reject a null hypothesis when the null hypothesis is actually true and it can be estimated by following Dwivedi et al.'s [73] simulation study (Algorithm 4). The two original samples come from the same distribution (with the same mean) and they are generated under the assumption that H_0 is true. For this simulation, the same choice of distribution parameters as in Dwivedi et al. [73] were adopted. The outcomes of the initial study for estimated type I error probability are presented in Table 2.2 (with nominal level $\alpha = 0.05$). Type I error above 5% is too permissive, whereas below 5% is too conservative. Permissive is generally considered more problematic than conservative.

Statistical power is the probability that the test correctly rejects the null hypothesis and it can be estimated by choosing two distributions with different means in Step 1 of Algorithm 4 and by replacing Type I error with estimated power in Step 6, otherwise the other steps of the simulation remain the same. The two original samples come from distributions with different means and they are generated under the assumption that H_1 is true. It is unclear from the article what parameters of the distributions were used for Normal and Skewed-Normal distribution, thus, those distribution parameters were chosen,

Algorithm 4 Estimating type I error for pairwise tests

1: Choose two distributions with the same mean;
2: Generate a sample from each distribution;
3: Perform the chosen pairwise test on those two distributions and record the p-value denoted by p;
4: In total, perform Steps 2 and 3 10,000 times to get $p_1, p_2, \ldots, p_{10,000}$;
5: Estimated type I error $= \frac{1}{10,000} \sum_{k=1}^{10,000} \mathbb{1}_{(p_k \leq \alpha)}$,

in a way that the difference between means is large. For Lognormal distribution, these parameters are provided by Dwivedi et al. [73]. The outcomes of the initial study for the estimated power are presented in Table 2.3. Higher estimated power is desirable.

The outcomes of this study are evaluated in relation to the two calculated metrics: estimated type I error and estimated power. As already stated, higher estimated power and estimated type I error close to the $\alpha = 0.05$ are desirable and conservative type I error ($\alpha < 0.05$) is more desirable than permissive type I error ($\alpha > 0.05$). This study concludes that in bootstrap hypothesis testing, Efron-B performs similarly to Banks-B finite, considering both the estimated power and type I error. Across most simulation cases, Banks-B finite yields slightly higher power than Efron-B, but also slightly higher type I error.

Using infinite range for Banks-B slightly decreases the type I error, but it also slightly decreases the estimated power. Similar pattern is found for NPI-B. For example, for sample size $n = 10$, for Normally distributed data from distributions with the same variance, Banks-B finite gives type I error 0.0521 whereas Banks-B infinite gives type I error 0.0492, the latter type I error (below $\alpha = 0.05$) is more desirable. On the other hand, for this case, the estimated power for Banks-B finite is 0.9908 whereas for Banks-B infinite it is 0.9899 and the former estimated power is more desirable. Therefore, it is up to decision maker to decide whether it is more important to reduce type I error or to increase power.

For most cases where the two original samples have unequal sample size, Banks-B yields higher estimated power than Efron-B, the Welsch t-test and the Students t-test. For Normally distributed data, the difference is even clearer for sample sizes 3 vs. 7 than

Distributions	Sample size	Efron-B	Banks-B finite	Banks-B infinite	NPI-B finite	NPI-B infinite	Student's t-test	Welsch t-test
N(5,1) v. N(5,1)	5	0.0505	0.0559	0.0524	0.0216	0.0203	0.0497	0.0436
N(5,1) v. N(5,1)	10	0.0468	0.0521	0.0492	0.0204	0.0175	0.0491	0.0475
N(5,1) v. N(5,1)	5 v. 10	0.0480	0.0578	0.0510	0.0252	0.0216	0.0493	0.0513
N(5,1) v. N(5,1)	3 v. 7	0.0422	0.0572	0.0485	0.0224	0.0198	0.0469	0.0555
N(5,1) v. N(5,9)	5	0.0701	0.0748	0.0723	0.0361	0.0348	0.0676	0.0543
N(5,1) v. N(5,9)	10	0.0580	0.0630	0.0586	0.0286	0.0257	0.0565	0.0473
LN_A v. LN_A	5	0.0425	0.0470	0.0439	0.0176	0.0167	0.0408	0.0303
LN_A v. LN_A	10	0.0449	0.0497	0.0455	0.0181	0.0153	0.0426	0.0365
LN_A v. LN_A	5 v. 10	0.0452	0.0548	0.0476	0.0223	0.0182	0.0435	0.0461
LN_A v. LN_A	3 v. 7	0.0370	0.0510	0.0414	0.0192	0.0154	0.0412	0.0492
SN(5,1) v. SN(5,1)	5	0.0500	0.0539	0.0515	0.0231	0.0201	0.0496	0.0438
SN(5,1) v. SN(5,1)	10	0.0471	0.0527	0.0496	0.0200	0.0188	0.0483	0.0467
SN(5,1) v. SN(5,1)	5 v. 10	0.0465	0.0563	0.0500	0.0253	0.0210	0.0490	0.0498
SN(5,1) v. SN(5,1)	3. v. 7	0.0425	0.0569	0.0552	0.0236	0.0204	0.0492	0.0552
SN(5,1) v. SN(5,9)	5	0.0734	0.0785	0.0734	0.0350	0.0334	0.0697	0.0556
SN(5,1) v. SN(5,9)	10	0.0606	0.0668	0.0619	0.0288	0.0256	0.0591	0.0512

Table 2.2: Type I error probability (nominal level = 0.05) for Efron-B t-test, NPI-B t-test (finite and infinite range), Banks-B t-test (finite and infinite range), Student's t-test and Welsch t-test

Distributions	Sample size	Efron-B	Banks-B finite	Banks-B infinite	NPI-B finite	NPI-B infinite	Student's t-test	Welsch t-test
N(5,1) v. N(7,1)	5	0.7913	0.8060	0.7943	0.6369	0.6161	0.7906	0.7679
N(5,1) v. N(7,1)	10	0.9899	0.9908	0.9899	0.9670	0.9621	0.9903	0.9897
N(5,1) v. N(7,1)	5 v. 10	0.8947	0.9125	0.9029	0.8182	0.7958	0.9220	0.8806
N(5,1) v. N(7,1)	3 v. 7	0.6396	0.7008	0.7943	0.5152	0.4801	0.1100	0.5878
N(5,1) v. N(7,9)	5	0.2733	0.2830	0.2735	0.1679	0.1588	0.2611	0.2090
N(5,1) v. N(7,9)	10	0.4887	0.5015	0.4900	0.3510	0.3346	0.4820	0.4466
N(5,1) v. N(7,9)	5 v. 10	0.3718	0.4134	0.3861	0.2666	0.2346	0.2208	0.4100
N(5,1) v. N(7,9)	3 v. 7	0.1818	0.2434	0.2062	0.1171	0.0939	0.1100	0.2648
N(5,1) v. N(6,1)	5	0.2863	0.3052	0.2926	0.1738	0.1617	0.2856	0.2652
N(5,1) v. N(6,1)	10	0.5624	0.5761	0.5636	0.4072	0.3858	0.5659	0.5595
LN_A v. LN_B	5	0.2909	0.2918	0.2755	0.1287	0.1176	0.2295	0.1377
LN_A v. LN_B	10	0.7010	0.6989	0.6873	0.4619	0.4349	0.6041	0.5240
LN_A v. LN_B	5 v. 10	0.5450	0.6015	0.5437	0.3494	0.2966	0.1737	0.5076
LN_A v. LN_B	3 v. 7	0.2231	0.3135	0.2493	0.1029	0.0717	0.0501	0.0133
LN_A v. LN_C	5	0.0452	0.0461	0.0421	0.0090	0.0079	0.0247	0.0072
LN_A v. LN_C	10	0.1301	0.1124	0.1085	0.0298	0.0243	0.0580	0.0328
LN_A v. LN_C	5 v. 10	0.0474	0.0666	0.0448	0.0123	0.0076	0.0018	0.0327
LN_A v. LN_C	3 v. 7	0.0070	0.0176	0.0082	0.0015	0.0008	0.0004	0.0133
SN(5,1) v. SN(7,1)	5	0.7873	0.8020	0.7914	0.6369	0.6145	0.7874	0.7655
SN(5,1) v. SN(7,1)	10	0.9867	0.9885	0.9630	0.9598	0.9871	0.9871	0.9863
SN(5,1) v. SN(7,1)	5 v. 10	0.9065	0.9262	0.9148	0.8307	0.8077	0.9242	0.8980
SN(5,1) v. SN(7,1)	3 v. 7	0.6194	0.6906	0.6512	0.4896	0.4462	0.7111	0.5738
SN(5,1) v. SN(7,9)	5	0.2937	0.3072	0.2963	0.1950	0.1853	0.2846	0.2390
SN(5,1) v. SN(7,9)	10	0.4931	0.5083	0.4952	0.3657	0.3480	0.4918	0.4587

Table 2.3: Empirical power for Efron-B t-test, NPI-B t-test (finite and infinite range), Banks-B t-test (finite and infinite range), Student's t-test and Welsch t-test

for 5 vs. 10. For sample sizes 5 vs. 10, for Normally distributed data where the two distributions only differ in means but not in variance, Banks-B finite gives estimated power of 0.9125, as opposed to Efron-B (0.8947), Banks-B infinite (0.9029), NPI-B finite (0.8182), NPI-B infinite (0.7958) and the Welsch t-test (0.8806). The Student's t-test still gives the highest power 0.9220. However if the original samples are of sample sizes 3 and 7, Banks-B infinite gives the highest power (0.7943), the second best power is given by Banks-B finite (0.7008). Efron-B performs worse (0.6396) and the Student's t-test gives the worst estimated power (0.1100). For Lognormally distributed data, for sample sizes 5 vs. 10, Banks-B finite gives the highest power (0.6015) compared to Efron-B (0.5450), Banks-B infinite (0.543), the Student's t-test (0.1737) and the Welsch t-test (0.5076). For Skewed-Normal distribution (with $Sk = 0.8$, two distributions of the same variance), the Student's t-test yields similar estimated power as Banks-B finite for sample sizes 5 vs. 10 and even better for 3 vs. 7. It would be of future interest to investigate Skewed-Normal distribution with larger skewing parameter.

For cases, where the two original samples come from two very different Lognormal distributions, such as LN(1,0.6^2) vs. LN(3,4^2), all tests have very low power, but bootstrap hypothesis test (Efron-B and Banks-B) perform better than traditional tests: the Student's t-test and the Welsch t-test. For example, for $n = 5$, Efron-B yields estimated power of 0.0452, Banks-B finite of 0.0461, Banks-B infinite of 0.0421, which is higher than for the Student's t-test (0.0247) and the Welsch t-test (0.0072).

Across all cases, NPI-B yields lower type I error than both Efron-B and Banks-B, and also lower estimated power. Neither very small type I error or low estimated power are desirable. For example, for sample size $n = 5$, for Normally distributed data from distributions with the same variance, NPI-B finite gives type I error 0.0216, NPI-B infinite gives type I error 0.0203, as opposed to Efron-B (0.0505), Banks-B finite (0.0559), Banks-B infinite (0.0524), the Student's t-test (0.0497) and the Welsch t-test (0.0436). But the estimated power for NPI-B finite is 0.6369 and for NPI-B infinite is 0.6161, as opposed to Efron-B (0.7913), Banks-B finite (0.8060), Banks-B infinite (0.7943), the Student's t-test (0.7906) and the Welsch t-test (0.7679). This can be explained by the fact that NPI-B creates more variability in both mean and variance of the t-statistic. Given the low power, NPI-B is not a suitable replacement for Efron-B.

As has been discussed throughout this chapter, the underlying distribution for small samples cannot be determined with certainty in real-life test scenarios, and, thus, a test comparing two samples should account for this. Bootstrap hypothesis testing is a plausible option for a two-sided test comparing two samples. The aim of this study was not to provide concrete guidance for practitioners, but to explore whether Banks-B could be used instead of Efron-B in bootstrap hypothesis testing. The initial findings show that Banks-B performs well, especially in cases of unequal sample size and variance of the two original samples that are being compared. However, NPI-B is not a recommended substitution for Efron-B in bootstrap hypothesis testing, as when it is applied, the method performs worse than it did with Banks-B or Efron-B, especially because NPI-B hypothesis testing leads to very low type I error probability, which is not desirable.

2.7 Concluding remarks

The aim of this chapter was to explore whether a bootstrap method can provide useful inference with small samples and to give initial recommendations on small-sample bootstrap to practitioners. Smoothened bootstrap methods, Banks-B and Hutson-B, showed a potential in the estimation of population characteristics for small samples in previous small-scale studies, however, these bootstrap methods are not as well known as Efron-B. NPI-B, Banks-B, Hutson-B and Efron-B were compared, focusing on their performance in the estimation and prediction of population characteristics for small samples. NPI-B, Banks-B, Hutson-B create bootstrap samples that contain observations that are different from the actual data, whereas Efron bootstrap samples with replacement from the original observations. NPI-B has been developed for prediction, but it was included in the comparison study of the bootstrap methods' performance in estimation to investigate how well it performs in estimation. For a similar reason, Efron-B, Banks-B and Hutson-B were included in the study of the bootstrap coverage performance in making prediction inference.

Summary of main findings

Findings of the simulation study encourage further exploration in the potential use of the bootstrap method for small samples. Banks-B performed well in the estimation of mean, variance and quantiles (Q1, median and Q3), regardless of the underlying distribution. Hutson-B showed good performance in the estimation of quantiles for a variety of underlying distributions. NPI-B proved to be the best performing bootstrap method in the prediction of mean, variance and quantiles. These conclusions are particularly promising in preclinical research, where sample sizes are small and the underlying distributions are not always known. Most real-life applications of the bootstrap method are from clinical research with large sample sizes, as opposed to preclinical research with small sample sizes, as discussed in Section 2.2. The main reason is that the most commonly known bootstrap method, Efron-B, is not regarded as a reliable method for very small sample sizes [42, 201] due to the method being based on the asymptotic argument, as described in Section 2.1.

Summary for each bootstrap method

NPI-B is the best performing method in the prediction of mean, variance and quantiles for Normally, Lognormally, Exponentially and Mixed-Normally distributed data for small sample sizes. Previous study by BinHimd [31] showed that NPI-B performs well in prediction for Normal, Uniform and Gamma distribution for sample sizes $n = 20$ and larger. Therefore, this study extends the findings of BinHimd to smaller samples sizes. The conclusion that NPI-B is good in the prediction for small sample sizes is beneficial to the second topic addressed in this thesis: statistical reproducibility. This topic is not limited to small samples but this thesis focuses on test scenarios with small samples in Chapter 4. The simulation study found out that NPI-B is not the best performing bootstrap in the prediction of IQR. More research needs to be done in order to explain the later phenomenon. Furthermore, NPI-B performs well in the estimation of population characteristics for sample size $n = 4$ and in the estimation of variance for Lognormally distributed data, although NPI-B was not created for estimation. In most cases, where NPI-B is used for estimation, there is over-coverage at 90% CI and this over-coverage

increases as n increases. This is due to the large variability of NPI-B samples.

Banks-B showed good performance in the estimation of mean, variance, median, Q1 and Q3 for small sample sizes ($n = 10$ and smaller) for a variety of distributions and this thesis recommends this bootstrap method for these scenarios. For data on the real-line, the recommendation is to use finite range (Approach I, Section 2.3.3) for Banks-B for the estimation of mean and quantiles and infinite range (Section 2.3.3) for the estimation of variance. For data which are equal or larger than 0, such as data with underlying Exponential or Lognormal distribution, half-infinite range (Approach V, Section 2.3.3) is recommended. Furthermore, Banks-B performs better than Efron-B in the prediction of mean, variance and quantiles for small sample sizes. However, this thesis would not recommend to substitute NPI-B by Banks-B for prediction. This chapter explored using Banks-B instead of Efron-B in bootstrap hypothesis testing (see Section 2.6) and the initial study showed that Banks-B performs well, in some cases even better, than Efron-B.

Efron-B showed good performance in estimation for large sample sizes (for the estimation of mean for Normal distribution from $n = 20$ and for the estimation of variance for Normal distribution from $n = 50$). Further exploration of Efron-B for large sample sizes is beyond the scope of this thesis. However, Efron-B performs poorly in the estimation of mean, variance, and quantiles for small sample sizes for data from both the Normal distribution and other distributions. It performs well in the prediction of IQR in some cases for small sample sizes, possibly due to the smaller variability of Efron-B bootstrap samples compared to the other studied bootstrap methods.

Hutson-B showed good performance in the estimation of variance for Normally distributed data and for Mixed-Normally distributed data (where the two distributions have different variances). However, Hutson-B performed poorly in the estimation of variance for Exponentially and Lognormally distributed data. Similarly, it performed poorly in the estimation of mean for Exponentially and Lognormally distributed data. Given that the underlying distribution cannot be ascertained for small samples, the work in this chapter would not lead to the recommendation of the use of Hutson-B for the estimation of mean and variance. Hutson-B showed good performance in estimating quantiles for both Normally, Exponentially and Lognormally distributed data, thus, further research into Hutson-B for the estimation of quantiles is encouraged. Hutson-B also showed good

performance in prediction for some cases, such as the prediction of Q1 for Lognormally distributed data of sample sizes $n = 4, 6$, the prediction of variance for Exponentially distributed data of size $n = 4$ and the prediction of variance of Normally distributed data of size $n = 4$. However, Hutson-B is not recommended for prediction as NPI-B is a more suitable and a better performing bootstrap method. Hutson-B has been defined only on the full infinite line $(-\infty, \infty)$ and on $[0, \infty)$. It would be of interest to consider developing Hutson-B for finite range, so that it can be applied to distributions defined on finite intervals.

Summary of the confidence interval choice

The simulation study briefly assessed whether using BC_a confidence intervals instead of percentile confidence interval would have an impact on small-sample bootstrap methods. It concluded that, from the perspective of either of the two metrics of assessment, for both Normally and Lognormally distributed data, Efron-B does not perform better than the other three bootstrap methods (Banks-B, Hutson-B and NPI-B) even when BC_a confidence intervals are used. There were cases where NPI-B, Banks-B (the estimation of quantiles for Normally distributed data) and Hutson-B (the estimation of mean and quantiles for Normally distributed data) performed better with BC_a confidence intervals. However, the findings of this study do not lead to the recommendation of using BC_a confidence intervals for the estimation of mean, variance and median for Lognormally distributed data for either of the four bootstrap methods. Given that the underlying distribution of the data cannot be ascertained, using percentile confidence intervals remains a safer choice. Practitioners should be careful about using BC_a confidence intervals instead of percentile confidence for the estimation of any population characteristics, before more research has been carried out. Nevertheless, this chapter encourages further research into the use of BC_a confidence intervals with NPI-B and Efron-B for the estimation of IQR, and with Hutson-B and Banks-B for the estimation of Q1 and Q3 for Lognormally distributed data.

Further research suggestions

The simulation study has opened up many questions and ideas for further research. It would be of practical interest to investigate further bootstrap method use with small sample sizes, possibly considering Banks-B as an alternative bootstrap method to Efron-B, especially where the population characteristics of interest are mean or variance. This investigation has been experimental, not theoretical, and the findings relate to the studied distributions. For applications beyond this study, it would be wise to run more comparison studies, relevant to the practical application circumstances. Future simulation study could extend this chapter's circumstances, by considering more cases for data from a variety of distributions, e.g. Laplace, Weibull, Beta, Uniform, Gamma distributions. Also, more cases of Mixed-Normal distributions could be investigated. Further investigation could consider the following properties: outliers, extreme skewness, a variety of mixtures, and kurtosis. The choice of distribution should reflect these relevant properties.

Section 2.4.3 pointed out that Hutson-B performs better in the estimation of Q1 than of Q3 for Lognormally distributed data. This could possible be caused by the influence of a heavy tail. Beta distribution could be used to study further the influence of the heavy tail on the performance of Hutson-B. Moreover, this exploration could also extend to other bootstrap methods and explore whether some bootstrap methods are affected more and some less by heavy tails. Moreover, Hutson [112] developed a sigmoidal quantile function estimator and a hybrid quantile function estimator, which was not enclosed in the study. It would be of interest of future research to study a variation of Hutson-B for small samples, using different quantile function estimator.

The choice of range for NPI-B and Banks-B has an effect on the bootstrap performance. Further exploration could extensively study the range choice. The findings regarding the bootstrap method performance in the estimation and prediction of IQR are inconclusive and a future study could carry out further investigation. Only percentile and BC_a confidence intervals were applied in the simulation study. Bootstrap-t confidence intervals are not recommended for small samples and in nonparametric situations [69]. However, both Banks [18] and Polansky [167] considered bootstrap-t confidence intervals in their comparisons studies, which involved small samples and Banks-B and Kernel-B, respectively. Thus, future study could include bootstrap-t confidence intervals.

The initial study of using Banks-B instead of Efron-B in bootstrap hypothesis testing (Section 2.6) showed that Banks-B performs well, in some cases even better, than Efron-B. It might be of interest to explore further whether Banks-B would perform better than Efron-B in cases where sample sizes of the original samples are unequal, and where data come from a variety of distributions. Two-sided t-test is commonly used in preclinical research, as a biomarker or a safety outcome variable could change in either direction. Thus, a future study could explore bootstrap hypothesis testing for real-life test scenarios where two-sided t-test is used.

Future research could also explore whether bootstrap hypothesis testing could be carried out under circumstances where the two samples are not combined, but rather bootstrap samples are generated from each original sample separately. This is not recommended for Efron-B, but it would be of interest to explore whether Banks-B could be employed instead of Efron-B.

Peng et al. [164] discussed the use of the bootstrap method in power and sample size calculations and provided examples from clinical research. These examples focused on large sample sizes ($n = 75$ and larger). This method could be further explored for small samples. A future study could also investigate how well would Banks-B perform if used instead of Efron-B for small samples. For in vivo studies with small sample sizes, power and sample size calculations are required to ensure that the experiment is appropriately designed to justify the use of animals and bootstrap method is a potential tool.

This chapter also touched upon some more distant research topics. In Appendix A.1, the issue of different types of quantile calculations leading to different sample statistics is addressed. The bootstrap methods' performance has been slightly affected by the type of quantile calculation. However, this thesis recommendations are not affected by the choice of the type of quantiles calculation. However, it would be of interest to explore what type is the most suitable for calculating sample quantiles for small samples. Similarly, bootstrap method requires that the data is representative of the population. *Population* stands for a complete set of individuals with a common characteristic and *sufficiently representative* means that meaningful analysis can be made for this sample. Linked to small-sample bootstrap, the following questions arise: Can small samples satisfy this criteria and, if so, how small can the sample size be to be sufficiently representative of the population?

Chapter 3

Reproducibility

3.1 Introduction

Reproducibility [12] is a complex issue, gaining importance and attention in scientific research. *Nature* published a special edition *Challenges in irreproducible research*, dedicated to the problem of researchers not being able to verify results presented in published papers of other scientists [151]. Ioannidis [118,119] drew attention to the high proportion of false research findings in the literature and he listed reproducibility practices, without defining them further, as research practices that can make more published research true [119]. In the existing literature on the topic of reproducibility there has been a lot of confusion about what the term reproducibility means [92], which will be addressed in this chapter.

A better understanding of reproducibility of tests is crucial for pharmaceutical research and development, as a lack of reproducibility contributes to failure rates in drug discovery and development processes, increasing costs, and decreasing efficiency. Begley and Ellis [23] highlighted a systematic problem in preclinical cancer research: the majority of publications in this research area cannot be validated. Scientists at the biotechnology firm Amgen tried to confirm findings of 53 published papers in haematology and oncology by performing replicate experiments. These did not reproduce conclusions in 47 out of 53 studies, even with the attempts to contact the original authors of the articles and to discuss the details of the experiments with them [23]. Errington et al. [81, 82] attempted to carry out replicate experiments based on high-impact papers published in 2010-2012 in the field of preclinical research in cancer biology. A replicate study is a new study, trying

to closely imitate the original study. Out of the chosen 193 experiments from 53 papers, they managed to conduct a replicate study for only 50 experiments from 23 original papers. 40% of replications of positive effects and 80% of replications of null effects were successful, according to three or more of five methods of replication assessment, defined by Errington et al. [82].

It is evident that scientists show considerable interest in the topic of reproducibility (or a lack thereof). The number of publications considering reproducibility is large. There is a rich body of literature on reproducibility in pharmaceutical research, particularly in preclinical research [34, 124, 130, 131, 175, 203], which will be discussed in Section 3.5. In psychology [133, 136, 157, 158, 195, 211], the focus is on the discussion of replicating the outcomes of the original study in a new replicate study and the concern about low reproducibility rate (or rather *replicability* rate, as this is commonly used in psychology). Computer sciences, machine learning and artificial intelligence [44, 94, 162] mainly focus on transparency and sharing of data, code and clear documentation of the whole study. Ioannidis [119] argued that sharing protocols, materials, software, and data provides a sound basis for reproducible data practices. This aspect is also important in chemistry [29, 90], nevertheless, there is also practical advice on how to maximise reproducibility through good laboratory practice and minimising human error.

The purpose of this chapter is to provide a review of the literature on reproducibility and highlight important debates on the topic. It is intended for a broad audience, including not only statisticians, but also anyone interested in the subject. By shedding light on the issue of reproducibility, this paper aims to contribute to the continuing discussion in this field. The aim of scientific research - to establish information about the nature of the world - is largely linked to reproducibility. Although reproducibility is part of the discussion on doing quality research, it does not equate to it. Thus, this chapter does not aim to address all aspects of quality research.

This chapter presents a literature review on the topic of reproducibility, which summarises several important debates. It is aimed at, but not restricted to, statisticians. Given that there are no standardised definitions for reproducibility and related terms (such as replicability), and that some definitions of reproducibility from the existing literature lack clarity themselves, this review begins by discussing various possible inter-

pretations and definitions of the concept of reproducibility in Section 3.2. Terms that are related to, or used interchangeably with, reproducibility are also discussed. With the aim to describe the subtleties encountered in the literature, the available definitions are classified into five categories, which we refer to as Type A to Type E. Section 3.3 briefly discusses goals of reproducibility provided in the literature, Section 3.4 outlines reasons for low reproducibility and suggestions for improvement of reproducibility presented in the literature, while Section 3.5 introduces some reproducibility issues related to preclinical research. Then Section 3.6 focuses on statistical reproducibility, classifying definitions of statistical reproducibility, and summarising important questions that have been raised. We also briefly comment on one of the debates in the reproducibility crisis discussion: whether p-values should be used.

This thesis focuses on statistical reproducibility. A majority of the literature is concerned with the validation of test conclusions, where both the original and the new (replicate) experiment have been carried out, this is addressed in Section 3.7. However, from the perspective of statistics, it is interesting to also study reproducibility in cases where only the original experiment has been carried out. Available methods for this approach to reproducibility will be introduced in Section 3.8. Section 3.9 defines and elaborates on the approach to statistical reproducibility adopted in this thesis: nonparametric predictive inference (NPI) reproducibility, placing the work in this thesis within the existing literature. Finally, Section 3.10 concludes and presents further research questions.

3.2 Definitions of reproducibility

There is no universally agreed definition for the concept of reproducibility and there are many related terms to reproducibility, such as repeatability, replicability, generalisability, robustness, reliability, open science, transparency, truth [152, p.36] and precision [116]. These related concepts are often also not clearly or appropriately defined, some of them are used interchangeably and they are all important for the reproducibility debate. This section presents a summary of definitions for reproducibility used in the existing literature.

Recent overviews of definitions of reproducibility and related terms have been presented by Goodman et al. [92], Barba [19] and Gundersen [94]. Goodman et al. [92]

identified that the term *research reproducibility* is not settled both linguistically and conceptually. Barba [19] raised the problem of different groups of researchers using different terminology for the same definition. The terms *reproducibility* and *replication* are often used interchangeably by researchers, which creates confusion and leads to conceptual ambiguity in the literature [19].

Rather than adhering to precise definitions, this work will classify the definitions that we encountered while reading the literature on reproducibility into five categories of reproducibility, Type A to Type E. The nuances are captured in the 'Reproducibility types tree' in Figure 3.1. This figure outlines possible considerations that are important for defining reproducibility and related terminology. In the descriptions of different types of reproducibility, three key terms, data, method and conclusion, are used. Data are *information, especially facts or numbers, collected to be examined and considered and used to help decision-making* [61]. The term method refers to the way the experiment is run. Method encompasses experimental design, data collection method, statistical analysis, software used to analyse the data and programming code. The range of features the method contains differs across different research areas. Conclusion is *a reasoned deduction or inference* [70], conclusion is reached after applying statistical analysis to the data. Next, the five Reproducibility types will be introduced.

Reproducibility Type A: Reproducibility is the ability to follow the analysis of an experiment based on the same data and a clear description of the data and the method.

Stronger version of **Reproducibility Type A:** Experimental conclusions are reproduced if another researcher applied the same analysis to the same data and reached the same conclusions, using the description of the data and the method provided by the original researcher.

Reproducibility Type B: Experimental conclusions are reproducible if same data but a different method of statistical analysis lead to the same conclusion.

Reproducibility Type C: Experimental conclusions are reproducible if new data from a new study carried out by the same team of scientists in the same laboratory, using the same method of experiment design and analysis, lead to the same conclusion.

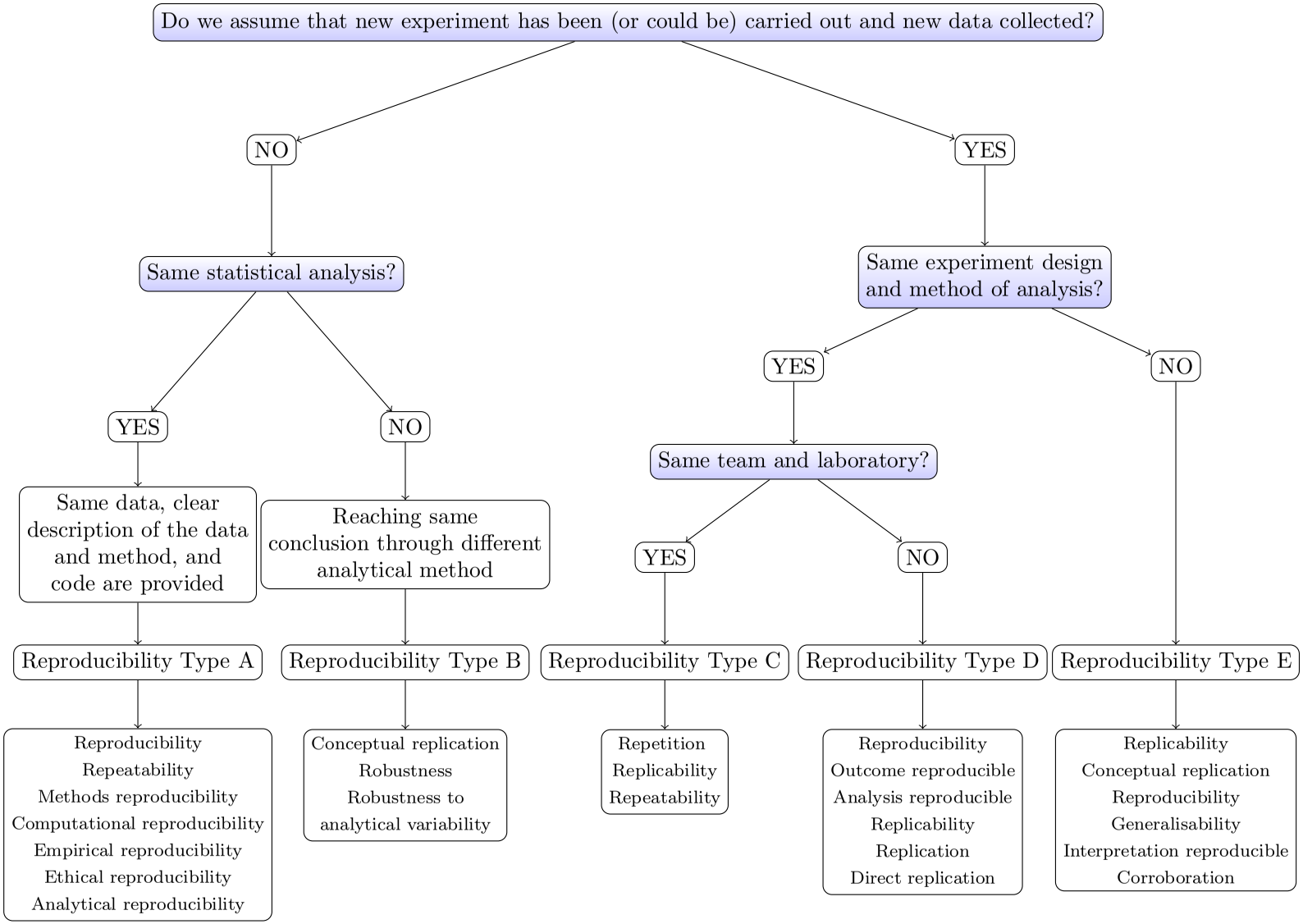


Figure 3.1: Reproducibility types tree

Reproducibility Type D: Experimental conclusions are reproducible if new data from a new study carried out by a different team of scientists in a different laboratory, using the same method of experiment design and analysis, lead to the same conclusion.

Reproducibility Type E: Experimental conclusions are reproducible if new data from a new study, using a different method of experiment design or analysis, lead to the same conclusion.

All the types of reproducibility, rely on a common underlying principle: namely, that the same conclusions 'would be' or have been reached in a reproducible experiment. Reproducibility Type A and Type B do not require new data, and Reproducibility Type C to Type E assume either the existence or the possibility of existence of new data. The term 'new study' is used in the Reproducibility Types instead of 'replicate study' because it is more general and it does not imply that the follow-up study exactly mimics the original study. This linguistic choice was made, as the term replicate study does not fit well Reproducibility Type E. Throughout this chapter, the terms new study and replicate study are used interchangeably.

There is a different debate, distinct from the Reproducibility Types classification, about whether there is a necessity to 'reproduce' (validate) the results by doing the experiment again. This debate deals with the question of whether the same conclusion 'would be' reached if the experiment was carried out again or whether the same conclusion has been reached after the new experiment has been carried out. Reproducibility Type C to Type E do not distinguish between these two options. Reproducibility has been assessed under both scenarios: First, when the new replicate experiment has been performed, which is addressed in Section 3.7. Secondly, when only the original experiment has been performed and probabilistic assessment is made about reproducibility based on the current data and analysis. Section 3.8 will focus on the latter scenario.

In the literature [11, 116, 125, 126, 133, 147], reproducibility has also been defined in terms of *precision* - the closeness of agreement between multiple (two or more) test results obtained under specific conditions, such as same method and same or different test operator. Blackman [32, 33] and Pryseley et al. [11] considered the quantification of this closeness of agreement between test results. In the above mentioned references, related

terms employed instead of *precision* are *reproducibility*, *repeatability*, *measurement precision*, *measurement repeatability* and *measurement reproducibility*. However, in these references there is no distinction between the original and the new test. The focus of these studies is on the variability in repeated measurements when one or more elements of the study, such as time, the observer, environment or instruments, are different [147]. We do not believe that *precision* should be considered equivalent to reproducibility. The concept of the original study is crucial to any discussion about reproducibility. Therefore, the assessment of the closeness of agreement is not within the scope of this overview. Note that the term *precision* used within this context is unrelated to the statistical concept of *imprecision* which was discussed in Chapter 1 and will be further discussed in Chapter 4.

Sections 3.2.1, 3.2.2 and 3.2.3 will elaborate on the variety of definitions for reproducibility and related terms available in the existing literature, and classify these terms into five different types of reproducibility. This classification into types aims to clarify different terminologies. It also aims to show inconsistency in terminology used across different publications and unclarity of some definitions. Each reproducibility type will be discussed separately, with the exception of Reproducibility Type C, Type D and Type E, which are discussed in the same section. The reason for this is that there are some definitions that either refer to multiple reproducibility types or it is unclear which of these three types they refer to. For each type, a list of terms used in the literature is presented.

3.2.1 Reproducibility Type A

Reproducibility Type A: Reproducibility is the ability to follow the analysis of an experiment based on the same data and a clear description of the data and the method.

Different terms, which classify as Type A:

- Reproducibility (National Science Foundation [155], Nosek et al. [154], Patit et al. [158], Nosek et al. [154])
- Repeatability (Proceedings of the National Academies of Sciences, Engineering, and Medicine [152])
- Methods reproducibility (Goodman et al. [92])
- Computational reproducibility (Donoho [72], Stodden [196])
- Empirical reproducibility (Stodden [196])
- Ethical reproducibility (Anderson et al. [7])

Stronger version of Reproducibility Type A: Experimental conclusions are reproduced if another researcher applied the same analysis to the same data and reached the same conclusions, using the description of the data and the method provided by the original researcher.

Different terms, which classify as the stronger version of Type A:

- Reproducibility (Benjamini, in the proceedings of NASEM [152, p.46], Stevens [195], Errington et al. [81])
- Computational reproducibility (National Academies of Sciences [153])
- Reproducible (Peng [161])
- Analytical reproducibility (Botvinik-Nezer and Wager [36])

In alignment to Reproducibility Type A, the requirement for reproducible research is that the documentation, data and code used for the analysis are available to others, so that they can verify the published results or carry out alternative analyses [161]. Goodman et al. [92] called *methods reproducibility* the ability, rather than necessity, to reach the same

conclusion by using the same data and method. Here ability refers to the availability of the data and a clear description of the data and the method, which would allow the researcher to re-enact the analysis. National Science Foundation's [155] definition of *reproducibility* can also be classified as Reproducibility Type A. Similarly, a workshop organised by the National Academies of Sciences, Engineering, and Medicine (NASEM) [152] used the term *repeatability* (also called *empirical reproducibility*), which can be classified as Reproducibility Type A. Donoho [72] used the term *computational reproducibility* without explicitly defining it, nevertheless its use is in alignment to Reproducibility Type A. Peng [162] argued that there is a spectrum of reproducibility. On the lower end of the spectrum is limited code sharing, in the middle section of the spectrum is sharing code and data, and on the upper end of the spectrum is sharing a single file containing both data and code that can execute the full analysis of the data. According to Peng [162], this upper end of the spectrum means full replication, which is the gold standard for reproducibility, as it allows the researcher to carry out the full analyses again [162]. Peng's spectrum of reproducibility does not encompass the term method, however, it is possible that this is because Peng discussed reproducibility in computational science, where code represents the method. Peng et al. [163] defined criteria for reproducible epidemiologic research as the availability of data, method, documentation and accessibility to the software, data, and documentation, which classifies as Reproducibility Type A. Gentleman and Lang [89] define *reproducible research* as research articles which are accompanied with software tools allowing readers to reproduce the paper results and further use the computational methods presented in the paper. Their definition can also classify as Reproducibility Type A.

Stodden [196] divided Reproducibility Type A into *empirical reproducibility*, which requires appropriate reporting standards and documentation of the physical experiment, and *computational reproducibility*, which requires accommodating the use of computation technology in the reporting and scientific practice. *Ethical reproducibility* [7], for which it is imperative to transparently report ethical challenges and methods of resolution of them in studies in biomedical research, also falls into the category of Reproducibility Type A. Thus there is a reasonable body of work that adopts definitions, which can be classified as Reproducibility Type A.

Reproducibility Type A leads to better transparency in research. We agree that care-

ful documentation of an experiment should be part of creating a reproducible research. We expect all research to have data, method, and code available upon request, but given the amount of the literature on definitions which classify as Reproducibility Type A, this is likely not the case. In computer sciences, Collberg et al. [44] conducted a study to determine whether 613 papers (from eight Association for Computing Machinery conferences and five computer science journals) presented reproducible research. Only papers, for which Collberg et al. [44] were able to obtain code and execute it, were labeled as reproducible research - reproducible in accordance with Reproducibility Type A. These were 102 out of 613. Collberg et al. [44] did not verify the accuracy of the published results. They provided an elaborate list of reasons why researchers did not provide code after email correspondence, examples of these reasons were: bad backup practices, the student who programmed the code left the research institution, and the code being an intellectual or commercial property.

A stronger version of Reproducibility Type A is presented by Benjamini in the proceedings of NASEM [152, p.46]. He defined *reproducibility* of the study as reaching the same conclusions after performing the same analysis on the study's raw data. *Reproducibility* in Errington et al. [81], Stevens [195] and Nosek et al. [154], and a consensus study report by the National Academies of Sciences [153] can also be classified as the stronger version of Reproducibility Type A. Botvinik-Nezer and Wager [36] called the stronger version of Reproducibility Type A *analytical reproducibility*. An article in *Biostatistics* is defined as reproducible if the Associate Editor for Reproducibility executed the code on the provided data and reproduced the results given in the article [161], which is also an example of the stronger version of Reproducibility Type A.

3.2.2 Reproducibility Type B

Reproducibility Type B: Experimental conclusions are reproducible if same data but a different statistical method of analysis lead to the same conclusion.

Different terms, which classify as Type B:

- Conceptual replication (Stahel [194])
- Robustness (Errington et al. [81])
- Robustness to analytical variability (Botvinik-Nezer and Wager [36])
- Arguably: Inferential reproducibility (Goodman et al. [92])

The core feature of Reproducibility Type B is that experimental conclusions are reproducible if the same data but a different method of data analysis were used to reach the same conclusion. While Reproducibility Type B is not a widely discussed kind of reproducibility, a reference to it can be found in Stahel [194], Goodman et al. [92], Errington et al. [81] and Botvinik-Nezer and Wager [36]. Stahel [194] discussed *conceptual replication*: where different analytical methods are used on the same data to re-examine conclusions of a study, which can be categorised as Reproducibility Type B. Errington et al. [81] used the term *robustness* for using alternative strategies on the same data, which also classifies as Reproducibility Type B. Similarly, Botvinik-Nezer and Wager's [36] terminology *robustness to analytical variability* fits with Reproducibility Type B.

Goodman et al. [92] presented *inferential reproducibility* which leads to similar conclusions from "an independent replication of a study or a re-analysis of the original study." The latter part of their definition could either refer to Reproducibility Type B or stronger version of Reproducibility Type A, depending on whether the re-analysis uses the same method as the original one did. The former part requires new data and new analysis, which is in alignment with Reproducibility Types C, D and E, which are discussed in Section 3.2.3.

It has not been clearly specified in the literature what is meant by 'different method of statistical analysis' in Reproducibility Type B, it is not clear how different this method can be and more thought should be given to this. Making sure that the statistical analysis is appropriate and suitable may be desirable. It is important to highlight that the negative

version of Reproducibility Type B, i.e. experimental conclusions being irreproducible due to different statistical reproducibility not leading to the same conclusion as the original statistical reproducibility, has not been considered in the literature. This negative version of Reproducibility Type B would be absurd, given that in many cases there is some statistical analysis that can lead to a different conclusion than the original statistical analysis. The reason behind the lack of exploration of these two mentioned aspects of Reproducibility Type B may be that reproducibility has been widely discussed by non-mathematicians and the discussion is lacking mathematical rigour.

3.2.3 Reproducibility Type C, Type D and Type E

Reproducibility Type C: Experimental conclusions are reproducible if new data from a new study carried out by the same team of scientists in the same laboratory, using the same method of experiment design and analysis, lead to the same conclusion.

Different terms, which classify as Type C:

- Repetition (Atmanspacher and Maasen [12])
- Replicability (National Science Foundation [155], National Academies of Sciences, Engineering and Medicine (NASEM) [153])
- Repeatability (Barba [19], Gundersen [94])

Reproducibility Type D: Experimental conclusions are reproducible if new data from a new study carried out by a different team of scientists in a different laboratory, using the same method of experiment design and analysis, lead to the same conclusion.

Different terms, which classify as Type D:

- Reproducibility (Voelkl et al. [202])
- Outcome reproducible (Gundersen [94])
- Analysis reproducible (Gundersen [94])
- Replicability (National Science Foundation [155], NASEM [153], Barba [19])
- Replication (Atmanspacher and Maasen [12])
- Direct replication (Zwaan et al. [211])

Reproducibility Type E: Experimental conclusions are reproducible if new data from a new study, using a different method of experiment design or analysis, lead to the same conclusion.

Different terms, which classify as Type E:

- Replicability (NASEM [153])
- Conceptual replication (Zwaan et al. [211])
- Reproducibility (Jarvis and Williams [124], Barba [19])
- Generalisability (National Science Foundation [155], Stahel [194])
- Interpretation reproducible (Gundersen [94]) – different analysis, same experimental design
- Corroboration (Gundersen [94]) – different experimental design

Definitions that encompass multiple Reproducibility Types:

- Replicable (National Science Foundation [155]) - encompasses Reproducibility Type C and D
- Replicability (NASEM [153], Errington et al. [81], Patil et al. [159], Stevens [195], Nosek et al. [154]) - encompasses Reproducibility Type C, D and E
- Reproducibility (Voelkl et al. [202], Richter [174]) - encompasses Reproducibility Type C and D
- Confirmation of conclusions (Stahel [194]) - encompasses Reproducibility Type C and D

Definitions for which it is unclear into which Reproducibility Type they fit:

- Inferential reproducibility (Goodman et al. [92]) – could be Reproducibility Type A, B, C, D or E
- Results reproducibility (Goodman et al. [92]) – could be Type C and D or just Type D
- Replication (Jarvis and Williams [124]) – definition compatible with Type C, possibly also Type D

A combination of Reproducibility Type C, Type D and Type E often fits to a particular definition. For example, a consensus study report by the National Academies of Sciences (NASEM) [153] defined *replicability* as "obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data." This rather broad definition of replicability includes Reproducibility Types C, D and E. The same definition of *replicability* was adopted by Errington et al. [81], Patil et al. [159] and Stevens [195]. Similary, Nosek et al. [154] refered to *replication* as using different data to test the reliability of prior finding.

In fact, it can be hard to distinguish under which type or types of reproducibility a particular definition can be categorised. Unclarity of definitions is a problem in the reproducibility debate. An example of an ambiguous and vague definition of reproducibility is Goodman's definition of reproducing the results of investigators in the proceedings of NASEM [152, p.41]. This is defined as "finding the same evidence or data, with the same

strength." It is unclear how to assess whether the requirement of this definition has been met.

Another example of unclear definitions has already been discussed in Section 3.2.2: *inferential reproducibility*. It is unclear what Goodman et al. [92] meant by the former part of the definition for *inferential reproducibility*: similar conclusions from "an independent replication of a study." It is unclear whether or not the circumstances of the original and the replicate study were identical or may have varied. If so, then it would classify as Reproducibility Type E. However, this is just a possible interpretation of the definition of *inferential reproducibility*. The definition could also fit with Reproducibility Type C or D. Moreover, this vague definition of reproducibility allows for the possibility of replicate study leading to considerably different data. In [152, p.42], Goodman defined *inferential reproducibility* as "reaching the same conclusions or inferences based on the results," however, this new definition does not yield more clarity.

Furthermore, Goodman et al. [92] defined *results reproducibility* as "obtaining the same results from the conduct of an independent study whose procedures are as closely matched to the original experiment as possible." It is unclear whether the definition categorises as both Reproducibility Types C and D or just the latter. The reason for this unclarity is that the term *independent study* has no clear definition and it is often used by researchers, including Goodman et al. [92], without being defined. It is unclear whether in Goodman et al.'s [92] definition of reproducibility, the same team of scientists or a different one has to carry out the experiment. On the other hand, Voelkl et al. [202] provided more clarity; they defined *reproducibility* as "the ability to produce similar results by an independent replicate experiment using the same methodology in the same or a different laboratory," which encompasses both Reproducibility Types C and D. A similar definition was used by Richter [174].

The National Science Foundation [155] distinguished three terms: *reproducibility*, *replicability* and *generalisability*, and they saw these as a foundation for robust scientific findings. *Reproducibility* can be categorised as the Reproducibility Type A definition, as stated in Section 3.2.1; *replicability* is the ability to validate the results of a prior study by collecting new data via the same procedure [155], which fits both Reproducibility Types C and D. *Generalisability* is attained when "the results of a study apply in other contexts

or populations that differ from the original one" [155]. Their definition of generalisability is most relevant to Reproducibility Type E.

Jarvis and Williams [124] defined *replication* as obtaining an identical result in an experiment conducted under identical conditions, which is compatible with Reproducibility Type C and possibly also with Reproducibility Type D, as it is unspecified whether the same team of scientists or the same laboratory is necessary. Jarvis and Williams [124] defined *reproducibility* as obtaining a similar result in an experiment conducted "under similar yet different conditions, the latter having the necessary degrees of latitude that reflect a real-world situation," which is most compatible with Reproducibility Type E. It is unclear what the terms 'similar results' and 'similar yet different conditions' exactly mean.

Barba [19] presented another division of terminology: *repeatability*, requiring the same team and the same experimental design, which can be categorised as Reproducibility Type C; *replicability*, requiring a different team and the same experimental setup and fitting in Reproducibility Type D; and *reproducibility*, requiring a different team and a different experimental design, which can be classified as Reproducibility Type E. It is consistent with definitions of *repetition* and *replication* by Atmanspacher and Maasen [12]: *repetition* refers to doing the same experiment by the same team whereas *replication* refers to situations where different teams carry out the same experiment. However, according to Atmanspacher and Maasen [12], *reproducibility* covers both terms. Gundersen's [94] use of the term *repeatability* also fits with Reproducibility Type C.

Zwaan et al. [211] defined *direct replication* as "studies intended to evaluate the ability of a particular method to produce the same results upon repetition." In this replication, critical elements of the study, such as procedures, samples and measures, are recreated [211]. But only "those elements that are believed necessary for producing the original effect" must be present in the replicate study. This definition is closest to Reproducibility Type D. Furthermore, Zwaan et al. [211] defined *conceptual replication*, which assesses whether an effect extends to a different population. *Conceptual replication* falls under Reproducibility Type E.

Gundersen [94] viewed *reproducibility* in the light of the scientific method. Gundersen's definition of reproducibility requires that a new experiment, mimicking the original

experiment by following the documentations from the original researcher, is carried out by another team of investigators [94]. Gundersen's categorisation of reproducibility differs from ours classification of reproducibility definitions. Gundersen [94] categorised reproducibility into four types of reproducibility, which define what type of documentation on the original study was available to the investigators carrying out the new study. Gundersen's four types are: description, code, data, and experiment. The last type encompasses all of the previously named three types. In our view, all details of the original study should be shared in order to replicate the study.

Furthermore, Gundersen [94] categorised reproducibility into three degrees of reproducibility, which he called *outcome reproducible*, *analysis reproducible* and *interpretation reproducible*. The degrees of reproducibility are based on what factors are similar in the original and the replicate experiment. *Outcome reproducible* means that the outcomes in the original and the replicate experiment are the same, thus applying the same analysis leads to the same conclusion. It is vague what outcome means in this context, nevertheless, Gundersen [94] stated that outcomes of some experiments are data. *Analysis reproducible* means that when the same analysis is applied to the new data in the replicate study, the same conclusion is reached. Arguably, both *outcome reproducible* and *analysis reproducible* fall under Reproducibility Type D. Interpretation of the analysis denotes the conclusion made about the study. *Interpretation reproducible* means that neither the outcome nor the analysis have to be the same in the replicate experiment, but the interpretation (conclusion) drawn from the original and the replicate study are the same. Thus, *interpretation reproducible* allows for different statistical analysis and it can be categorised as Reproducibility Type E.

Gundersen [94] argued that using different methodology no longer falls under reproducibility, but it is called *corroboration*, as in such cases hypotheses are supported by new evidence. *Corroboration* fits the best within Reproducibility Type E. Gundersen [94] also stated that corroboration refers to theories and hypotheses rather than to experiments. This underlines that there is a considerable disagreement on what constitutes reproducibility.

According to Stahel [194], there are two aspects of a successful replication. One of the two aspects is the *confirmation of conclusions*, which means that a replication study leads

to the same conclusion as the original one. This is compatible with Reproducibility Types C and D. Stahel [194] also noted that the concept of reproducibility can be extended to exploring different circumstances and if the new study leads to the same conclusion as the original study, then this is called *generalisability*, which could be interpreted as using a different method from the original study, and can be classified as Reproducibility Type E. Another aspect of a successful replication, according to Stahel [194], is *statistical compatibility*, which addresses the question "Is the data obtained in the replication compatible with the data from the original study in the light of the model used to draw inference?" This approach to reproducibility is not discussed elsewhere in the existing literature and this thesis does not classify this as any of the Reproducibility Types because it is not clear what does 'data from a new study are compatible with the data from the original study' exactly mean.

3.2.4 Summary

There is no universally accepted definition of the term *reproducibility*. There are many related terms, such as *repeatability*, *replicability*, and *generalisability*, which are also not clearly defined. Sometimes the same terms with different definitions are used across the literature. Moreover, researchers at times use the word reproducibility without defining the term, and even if they do so, the definitions are not always clear. Terms used in some of the definitions of reproducibility are not clearly defined. In order to have meaningful debates on reproducibility, it is important to clarify terminology relating to reproducibility. In this section, five types of reproducibility were identified and definitions from the literature, which can be classified as these types of reproducibility, were mentioned. Figure 3.1 outlined various considerations that are important in defining reproducibility and how these relate to the five types of reproducibility. There are some overlaps between the presented five types of reproducibility. It is not the purpose of this thesis to study which of these types of reproducibility is the 'correct' one. Arguably, all of the considerations presented in the five types are relevant to reproducibility of scientific findings.

3.3 Goals of reproducibility

In the discussion of reproducibility, it is important to consider what the goals of reproducibility (or of reproducibility assessment) are. A clear goal is confirmation of conclusions. But researchers do not have to limit themselves to this goal. According to Bayarri and Mayoral [22], other goals of reproducibility include: reduction of random error, bias detection and extension of conclusions. The latter goal relates to Reproducibility Type E. Goodman [152, p.42] extended the list by adding two additional goals: learning about the *robustness* ("resistance to minor or moderate changes in the experimental or analytic procedures and assumptions" [152, p.42]) and the *generalisability* of results ("true findings outside the experimental frame or in a not-yet-tested situation" [152, p.42]). Arguably, the *robustness* and *generalisability* are not additional goals, but rather clarifications of the fourth goal, extension of conclusions. Zwaan et al. [211] introduced another role of reproducibility, which they called *replication*, to provide more accurate estimates of effect sizes. This goal is of a different nature than the rest of the goals and it is questionable whether reproducibility should focus on estimation of effect sizes or whether effect sizes should be part of the general statistical analysis discussion. Further consideration of this is outside the scope of this thesis.

3.4 Reasons for low reproducibility and suggestions for improvement

This section provides an overview of reasons for low reproducibility and suggestions for improvement, based on the literature. Section 3.4.1 considers the topic from a statistical perspective, and Section 3.4.2 provides more general insights. Note that this chapter refers to a variety of literature, where authors often assume different definitions of reproducibility, as discussed in Section 3.2. Sometimes, a definition of reproducibility is assumed and not explicitly stated.

There is no universally accepted notion of what low or poor reproducibility means, which is likely linked to the lack of a universal definition for the term *reproducibility*. There are two main approaches to defining low or poor reproducibility: First, it can refer

to a poorly described and documented experiment, which prevents another researcher from reproducing the original experiment. This could be done either by using the same data, in alignment to Reproducibility Type A or Type B, or by redoing the experiment and acquiring new data, in alignment to Reproducibility Types C to E. Secondly, low reproducibility can refer to a well described and documented experiment, where a new experiment does not lead to the same findings that were reached in the original experiment. Poor reproducibility can also refer to a combination of these two approaches.

The solutions for improving reproducibility often require adhering to good scientific practice and using appropriate statistical, experimental and documentation methods. The majority of these solutions are not limited to a particular type of reproducibility. Finding solutions to the reproducibility crisis calls for many stakeholders: researchers; institutions, both public, such as universities, and private, such as companies; funding bodies; and journals. All these stakeholders can play a vital role in improving reproducibility [152, p.21]. Poor reproducibility may not be in the interest of any company.

3.4.1 From the perspective of statistics

Poor statistical choices

The discussion of statistical reasons for poor reproducibility begins by highlighting the problem of researchers making poor statistical choices. Wrong or unsuitable statistical analysis [16, 28, 86, 153, 171, 196] and poor experimental design [16, 153] are commonly named. This includes the incorrect use of p-values [28], overrelying on p-values [99], inadequate sample sizes [86,171], low power [196], using inappropriate sampling techniques [196], insufficient knowledge of data-generation mechanisms caused by the use of big data [196], experimental biases [28], statistical biases such as confouding [28], and programming errors [28]. The discussion of the reasons for low reproducibility in the quoted papers is mostly theoretical and it does not include real world examples.

More specific reasons for low reproducibility, which only apply to certain experiments, are: examination of weak and complex interactions for data with low signal-to-noise ratio [161], and miscalculation of effect sizes in meta-analyses [3]. Moreover, greater availability of data and more complicated analytical methods lead to a greater risk of

false or misleading findings [161], as this increases the risk of an error.

Statistical solutions to problems offered in the literature are: using suitable statistical methods [16, 28, 153], which may include reporting confidence intervals rather than just p-values [99]; using robust designs [16]; acknowledging uncertainties [153] and taking into consideration the sensitivity of estimates for both deviations in the underlying data and model choice [196]. To ensure the appropriate use of statistics, it is important to involve statisticians at all the stages of the experiments or to provide good statistical training to the researchers carrying out the experiments [136,173]. It is important to teach researchers that statistics is a tool to assess the strength of evidence, rather than to reveal the truth. Even, with this priority, occasional human error is still inevitable.

Berger [28] suggested using Bayesian analysis and he argued that this statistical framework provides a systematic way of dealing with multiple statistical analyses. Researchers not limiting themselves to either frequentist or Bayesian statistics, is desirable, however, this again requires proper statistical training. Stahel [194] encouraged cross validation; in cross validation, the dataset is split many times into a small training set and model parameters are estimated for each of these training sets, then the average performance of all splits is calculated. Using appropriate statistical analysis has the potential to reduce the incidence of wrong conclusions, which are often caused by technical errors.

Undesired correlations

Apart from the incorrect use of statistical methods, unwanted or unknown correlations could negatively affect reproducibility. Stahel [194] named the within laboratory or within group correlation, which is about measurements from the same laboratory being more similar. Stahel [194] suggested that there may be a correlation between results obtained with short time lags. Carrying out the same experiment in a different laboratory and by a different team of scientists, in line with Reproducibility Type D, can ensure that the conclusions of the experiment are not linked to some of the unknown correlations, such as the within laboratory correlation, which can occur with Reproducibility Type C. One way to reduce the undesired consequences of unwanted correlations is accounting for reproducibility in the design of the experiment. This solution has already been addressed in preclinical research, and this topic will be further discussed in Section 3.5. Repro-

ducibility Type E avoids some of the pitfalls of unwanted correlations, namely, it tests the findings of the experiment under changed circumstances, which makes the conclusions more robust with respect to the varying conditions. According to Ehm [79], meta-analysis is needed because of the issues of heterogeneity and selection bias [79]. Meta-analysis is a statistical method that combines results of several independent studies [95]. It should not replace replication studies, but it is useful as it can stop researchers from prematurely accepting conclusions, and from performing ineffective or harmful treatments.

Within-study selection bias

Related to the undesired correlation is the within-study selection bias. Hutton and Williamson [113] showed, via a meta-analysis on a treatment for incontinence and anthelminth therapy, that selective reporting of outcomes - when less outcomes are reported than are measured - has an effect on the conclusions and recommendations made about treatment. This within-study selection bias is often based on the significance level and the estimates of effect size. However, this selective reporting becomes problematic when meta-analysis is carried out or someone else tries to replicate the experiment. To avoid these problems, all outcomes should be reported, even those that were statistically insignificant.

Missing data

In research, missing data and the lack of documentation of missing data [189] can lead to poor reproducibility. Thus, it is important to report information about missing data, which includes the degree of and statistical assumptions related to missing data, and the practical information on reasons behind missing data. Moreover, it is vital to perform sensitivity analysis to assess robustness of these assumptions in order to increase reproducibility [189]. This solution seems feasible as the treatment of missing values is a part of the statistical analysis, and reporting them is in alignment with Reproducibility Type A.

Multiplicity

Multiplicity [92] or failure to adjust for multiplicities [28] can also lead to lower reproducibility. Multiplicity occurs when several statistical inferences are considered simultaneously, this often involves using multiple statistical tests. According to Bretz and

Westfall [37], ignoring multiplicity in any stage of drug development may cause a lack of reproducibility, which they call *replicability*, at a later stage or after market approval. Bretz and Westfall [37] carried out a simulation of pairs of independent studies, S_I and S_{II}, which only differ in the sample size: $n_{S_I} = 100$, $n_{S_{II}} = 1000$, where S_I represents the original test study, and S_{II} represents the replicate study. Everything else remains the same. This relates to Reproducibility Type E, with only one change of circumstances: the sample size. Bretz and Westfall [37] recorded the observed effect sizes of the selected populations for each study and they compared them. They concluded that the effect sizes of S_I are not 'reproduced' in S_{II}: on average they are larger than effect sizes of S_{II}. This confirms that changing one aspect of the test, such as the effect size, can have an effect on the test conclusion.

Deliberate statistical malpractices

Intentional statistical malpractices are another cause of poor reproducibility. These include: removing 'outliers' and unfavourable data [28], trying out multiple models until one gets favourable results [28] (also called p-hacking [92, 153] or selective reporting [16, 92]), statistical overfitting [14], data dredging (analysing data in order to find any possible relationships between the data) [92] and hypothesizing after the results are known [92, 153]. Such malpractices often stem from the pressure to publish [16]. Clear documentation of all statistical processes, which links to Reproducibility Type A, allows an external scientist to check the analysis carried out and it increases the chance of spotting statistical malpractices. Moreover, pre-registration of studies is conducive to transparency [36, 81, 150, 153, 193] and it prevents many malpractices. In pre-registration of studies, experimental designs and analytical plans are written down in a database before the experiment is performed. In clinical studies, pre-registration is mandatory and can be done through registries such as the International Standard Randomized Controlled Trial Number registry [117] and the International Clinical Trials Registry Platform [115].

3.4.2 More general insights

The majority of reasons for low reproducibility do not stem from wrong statistical analysis. Preference of publishing positive results, a lack of documentation of experiments, focus on exploratory studies rather than replication studies, and other non-statistical issues can lead to low reproducibility. This section will summarise these problems and outline suggestions for improvement offered in the literature.

Preference of publishing positive results

One of the reasons for low reproducibility is the pressure to publish [16]. This is exacerbated by the publication bias [28, 153, 194, 211], which refers to the preference of journals to publish positive results and reject negative results [86]. Similarly, negative or null results are also often not written up for publication. This leads to a high proportion of 'false positive' results.

Journals should strive to accept for publication articles with negative or null results [23, 28, 45]. Removing stigma associated with negative results, i.e. negative perception of negative results, has potential to increase reproducibility [198]. However, even if journals allow the publication of negative results, it is questionable whether researchers will start writing up negative results, simply because they are focused on positive results and they face many pressures, which prevent them from writing the negative results for journal publication. It is also questionable whether scientists would worry about reproducibility of negative results. If not, false negatives would be more problematic for science than false positives because they would receive less attention and scrutiny. An investigation carried out in this PhD project, that lead to null findings, is briefly described in Appendix B.3.

Other problems in the publication system

Allison [3] emphasised that there is a lack of formal guidance for post-publication corrections. He pointed out that in science a degree of self-correction is crucial, however, it is hard to be achieved via publications. Once an article gets published, it is hard to address any errors. The US National Institute of Health (NIH) promoted that journals should be motivated to allocate more space for papers that point out errors in earlier work [45],

which seems a feasible and forward-looking solution. This policy is being adopted by many journals.

Other problems in the publication system, that lead to lower reproducibility, are fraudulent research [16, 153, 161], insufficient peer review, oversight and mentoring [16] and competition between laboratories leading to hastily written papers [86].

Documentation

Incomplete or bad reporting [86,92,196] or a lack of 'instruction material' for scientists who want to produce such reproducible research [81, 161] can also lead to low reproducibility. The raw data, method description or code are not always available [16]. One of the reasons why scientists may not share their data and code is that it takes a lot of time to document the work and to clean up the data and code [138]. Alternatively, the researchers may not want to share documentation which includes work beyond the published results.

Iqbal et al. [122] carried out a systematic assessment of the biomedical literature, assessing transparency and reproducibility in a random sample of 441 articles in biomedical journals published in 2000-2014. They concluded that the biomedical literature lacks transparency; it is missing protocols, data, statements of conflict, funding information, and statements of novelty or replication. Similarly, Errington et al. [81] highlighted problems for replication of experiments: incomplete documentation, not enough information to repeat an experiment, descriptive or inferential statistics not provided, insufficient detail about the experiments. In their reproducibility project in cancer biology, the team of scientists sought to repeat 193 experiments from 53 papers [81]. In order to do so, they had to modify 67% of the protocols, which were already peer-reviewed, and they were only able to implement 41% of those modifications [81]. Only 4 out of 193 experiments included data which were necessary for computing the effect size and for conducting power analysis [81]. Following difficulties faced during the design and conduct of the experiments, Errington et al. [81] were only able to carry out new experiments based on the original experiments for 50 experiments from 23 papers.

Many sources agree that careful documentation of all steps in an experiment is important for reproducible research [14, 17, 28, 87, 89, 153]. From the perspective of the five reproducibility types, clear documentation is important. This includes code, data and

clear description of the data and the analysis [38, 180]. Moreover, public access to all these documents is necessary, so that other researchers can validate the analysis [179]. Berger [28] suggested establishing protocols for scientific investigations. Donoho [72] advised to create a single R script that generates all the results, figures and tables, for a particular paper. Solutions in terms of user-friendly software are not new; in 2000, Schwab et al. [180] described ReDoc, a simple software system where authors deposit all the documentation, data, and code, that allows readers to reproduce computational results from the articles. A more recent tool for a clear documentation of the statistical analysis is the R package *knitr* which creates a single document containing both the code and the documentation of the experiment, including visualisations [209]. Another solution is the use of a Jupyter notebook. This is a web-based computational environment that shows the code for data analysis alongside text and visualisations [179]. All these solutions are compatible with the approach to reproducibility described in Reproducibility Type A, making it possible for another researcher to go through the data, code, and the method, and reanalyse the experiment. However, these solutions are not limited to Reproducibility Type A; they are useful for researchers who want to analyse the same data using a different analytical method (Reproducibility Type B), or for researchers who want to repeat the experiment (in line with Reproducibility Type C, D or E). A difficulty is that these solutions are time-consuming and require training. Nevertheless, the long-term benefits of these solutions are apparent and they have already been implemented, for example, in computer sciences. A related question is what to do when irreproducibility is reported, in line with Reproducibility Type A or its stronger version. However, this question remains outside the scope of this thesis.

The recommendations on documentation should not be limited to technical aspects, such as what software to use, but they also should include a discussion of what information should be included and in what depth. A detailed manual about journal reporting in quantitative research in psychology can be found in Appelbaum et al. [9]. The recommendations presented in this article can also be applied to other research fields. An earlier work includes Wilkinson et al. [207] who give a detailed and useful guide for practitioners in psychology on how to carry out appropriate statistical methods, devise a good experimental design, document the work well. This article does not limit itself to the field

of psychology. Reporting guidelines for a broad spectrum of health research studies are given by the Enhancing the Quality and Transparency Of health Research (EQUATOR) network [80].

Tiwari et al. [199] proposed a 'reproducibility scorecard' for publications to improve reproducibility. This scorecard asks 8 questions, two examples of these questions are: "Are the model codes deposited in a relevant open model database?" and "Are the mathematical expressions described in the manuscript or supplementary material?" [199]. Tiwari et al. suggested a 4 out of 8 cut-off in the reproducibility scorecard. This means that 'yes' needs to be answered to at least four questions for there being a chance of reproducing the same results. Tiwari et al. [199] limited the scope of their paper to systems biology modelling but this idea could also be used in other scientific areas. However, rather than using a cut-off-point, it might be better to report which 'reproducibility criteria' the publication satisfies. These reproducibility criteria could encompass all five reproducibility types.

In big data settings, keeping track of all steps in an experiment and data management becomes a challenge. There are many tools that make it possible for researchers to document their work, such as an open-source programming language, a cloud-based data repository, a programming interface and the previously mentioned Jupyter Notebook [183]. Moreover, following the FAIR (Findable, Accessible, Interoperable and Reproducible) principles [208] for data management can lead to higher reproducibility [179]. FAIR principles are guidelines that guide researchers on how to organise, describe, store and operate data in order to improve the reusability of data. See Wilkinson et al. [208] for a more elaborate description of these principles.

Cooperation

Within an organisation, better reproducibility, in alignment with all reproducibility types, can be achieved through collaboration of a team [150] and the inclusion of statisticians in research teams [28]. This is linked to the need for interdisciplinary teams on large scale projects [179] and for initiative to share common vocabulary. This would allow for more informed conclusions. Moreover, better mentoring and supervision, better teaching, more within-lab validation, incentives for diligent work, and more external-lab validation can

improve reproducibility [16].

Focus on replication studies

Funding bodies also have a role to play as they can have an impact on reproducibility through grant distribution. There should be distinguishment between exploratory versus confirmatory studies [202]. Pusztai et al. [172] proposed that some of the existing funding from new-discovery oriented grants gets allocated to confirmatory and validation grants that could be used for verification of important published results. For example, Iorns et al. [121] presented a successful replication study in biology. The details of the experiment are omitted as these include biology related terminology. Iorns et al. [121] communicated with the original authors to receive further details of the study and they also performed additional analysis to collect more detailed data, including data for more doses than the original test scenario. This replicate study confirmed the conclusions of the original test scenario. However, the effects seen in the replicate study were lower than in the original study. To improve reproducibility, such efforts should receive recognition in the scientific community. However, such recognition might be hard to establish, as more emphasis is given to the discovery research and to the publication of novel findings. The Science Exchange network [181] established a support network for researchers who want to carry out replication studies in order to validate key experimental findings.

3.5 Reproducibility in preclinical research

As this thesis focuses on reproducibility in preclinical research, this section will address some of the issues regarding reproducibility of studies that specifically relate to this field. The motivating test scenarios in this thesis come from preclinical in-vivo research, i.e. research carried out on animals, typically rodents. Preclinical research mostly focuses on the actual replication of an experiment in accordance to Reproducibility Type E, as due to the inevitable variations between experiments, it is impossible to have exactly the same conditions in two separate experiments. Arguably, this is impossible in any area. Quantifying reproducibility, in situations when only the original experiment has been carried out, has not received much attention in preclinical research.

Section 3.5.1 considers ethical issues that limit preclinical research and result in the use of small samples. Section 3.5.2 presents other reproducibility challenges for preclinical research. Section 3.5.3 presents some recommendations for improving reproducibility in preclinical research. Section 3.5.4 is dedicated to the topic of embracing variability.

3.5.1 Ethical issues

Animals are a fundamental part of preclinical research and the majority of the discussion on reproducibility in preclinical research is linked to them. Due to ethical issues, sample sizes in animal studies are small. Thus, poor reproducibility may be to some extent unavoidable. On the other hand, a follow-up study, which assesses reproducibility, increases the number of animals needed [174]. The 3Rs principles [84] provide guidance for researchers on how to responsibly conduct experiments in animal research. The 3Rs stand for replace - animals by non-sentient animals whenever possible; reduce - the number of animals; and refine - improve animal well-being. The 'reduce' principle is the most relevant one in the discussion on reproducibility and there is, arguably, a need for a move from the traditional focus on reducing the number of animals per experiment solely to a more integrated approach which considers validity, robustness and reproducibility of experiments. The 'replace' and 'refine' principles are indirectly linked to the reproducibility debate: the more a researcher adheres to these principles, the more ethical ground there will be to repeat the experiment or to use a larger sample size. For the 'reduce' principle, an important question arises: Is it possible to improve reproducibility using smaller sample sizes, thus reducing the number of animal, assuming the experiment is set up optimally? This question is outside the scope of this thesis, however, it is important for future research.

3.5.2 Challenges using animal in research

Small sample size, linked to the ethical concerns, as well as to financial and practical reasons, is only one of the challenges a researcher faces when working with animals in preclinical research. The involvement of animals adds additional uncontrollable variability. Animals are very perceptive to small environmental changes, such as light and noise

coming out of a computer, and this can have an impact on the experiment.

Apart from the variations related to animal use, experiments may face the problem of inevitable variations, such as time lag, variation of apparatus and material [194]. Similarly, variability of standard reagents [16] can affect the experimental outcomes. Slightly changing the experimental procedure or using different laboratories, or different animal strains are some of the reasons for low reproducibility of experiment [45]. Here strain stands for a group of animals that are genetically the same.

Stevens [195] named other reasons, with focus on animal use in comparative psychology: There is often repeated testing on more animals that are more expensive than rodents, such as parrots or primates. Also people may have more objections to testing on more intelligent animals. Therefore, as much data as possible are collected during one experiment. This exploratory data analysis may lead to data fishing. Furthermore, there is often limited species coverage and species are often substituted in a replicate study.

3.5.3 Recommendations offered in literature

This section will discuss the advice offered in the literature on how to improve statistical training, planing and experimental design, and documentation of experiments in order to improve reproducibility in preclinical research. Embracing variability, a highly discussed recommendation, will be addressed in Section 3.5.4.

Reynolds [173] pointed out the lack of adequate statistical training in preclinical research and he advocated training in statistics for researchers, specific to preclinical research. According to Reynolds [173], researchers should be taught to create the statistical design and carry out data sampling, before analysing the data and making inferences. The importance of statistical training has already been discussed in Section 3.4. However, not much attention has been paid to the details of such statistical training, possibly because a lot of the literature has been written by non-statisticians. It would be desirable to discuss in greater depth the methods that should be taught, the level of understanding of the methods that researchers should acquire, and the guidance on when a non-statistician should consult a statistician.

It is important to note that an institution needs a license to carry out experiments on animals, and, in order to obtain this licence, the institution needs to show that it

provides good training. The UK home office provides the license but it takes an input from the Animal Welfare and Ethical Review Bodies (AWERBs). The AWERBs focus on high standard of animal welfare, improving scientific quality, promoting a culture of care and ensuring the 3Rs principles are followed [178]. One of the factors of maintaining a high quality of the above mentioned is training and this includes training in statistics and experimental design. "The AWERB needs to be confident that the establishment has in place a good system of education and training and assessment of competence for all staff [178]." However, there are no universal requirements on the specifics of the training in statistics.

Spanagel [193] recommended a variety of measures that can be incorporated into the planning and design of an experiment in order to improve reproducibility: Prior to a new study, researchers could consider conducting a systematic review or potential meta-analyses of existing related studies, conduct a power analysis, pre-register experimental study protocols, as discussed in Section 3.4, and consider carrying out multi-centre pre-clinical studies. In the context of research on psychiatric disorders, Spanagel [193] advised researchers to consider using animal models that satisfy two psychiatric diagnostic classification systems which are based on observations from clinical research [193], and it is important that the preclinical study reflects those. It is also advisable not to overcomplicate statistical analysis and to use only the methodology that the researcher has a good understanding of [193]. Richter [174] argued that the risk of bias could be prevented by random treatment allocation, blind administration of the treatment, and blind assessment of outcome. According to him, this could eliminate aspects of the experiment which lead to misleading results. However, it is arguable whether randomisation is preferable to careful balancing an experiment with known factors.

Regarding the documentation of an experiment, diligently following ARRIVE (Animal Research: Reporting of In Vivo Experiments) guidelines [10] improves reporting standards in animal testing [174] and thus makes replication of the experiments easier. ARRIVE guidelines provide directions on reporting of ten essential items: study design, sample size, inclusion and exclusion criteria, randomisation, blinding, outcome measure, statistical methods, experimental animals, experimental procedures, and results. Moreover, reproducibility can be improved by making raw data available in accordance to FAIR

principles [208] and by publishing negative findings [193], both recommendations have already been discussed in Section 3.4.

3.5.4 Heterogenisation – embracing variability

In alignment with Reproducibility Type E, there is a body of literature suggesting that systematic heterogenisation rather than standardisation improves reproducibility in preclinical research [34, 130, 131, 174, 175, 202]. This literature focuses on experiments carried out on mice. Richter [174] argued that perfect homogenisation decreases inter-individual variation within a study population to zero, which leads to statistically significant results that cannot be generalised to slightly different conditions. This is also called the *standardisation fallacy*. Standardisation does not account for animals being responsive to the environment, also known as phenotypic plasticity [130]. This biological variation caused by phenotypic plasticity differs from random noise [202]. In preclinical research, it has been suggested to embrace variability through systematic heterogenisation in order to improve reproducibility [130].

Examples of heterogenisation named in the literature are using mice of diverse characteristics, such as mice of different age, sex and body weight, [176]; using different inbred strains of mice [202]; co-housing individuals of different strains of mice [174]; varying the housing conditions of mice [202]; varying husbandry and test procedures [176]; and carrying out the experiment on mice at different times [34] or in multiple laboratories [203]. For example, Bodden et al. [34] presented a study where systematic heterogenisation, adding variability, via carrying the experiment on mice at different times of the day improves reproducibility (Type E).

A possible tool for heterogenisation is the use of randomised block designs for the experiments. This can include using time or a batch as blocking factors [85, 131]. The latter is called the multi-batch design where the experiments are split into small batches of animals which are tested at different times. These 'mini-experiments' are then brought together in the statistical analysis. Karp et al. [131] showed how multi-batch design improves reproducibility in a syngeneic tumour case study. For the multi-batch design, they explored the following statistical analyses: meta-analysis, a fixed effect regression approach, a random effect regression approach and a pooled approach [131]. A pooled

approach was not recommended for the statistical analysis as it ignores batch information. Meta-analysis and random effect regression were recommended by the authors for the analyses of multi-batch design experiments [131].

Embracing variability also addresses a problem that is interlinked with reproducibility: there is a high failure rate in translating research from preclinical to clinical studies [174]. Translating research means that conclusions about a new treatment reached in the preclinical stage of the drug development are validated in clinical research [174]. In a pharmaceutical context, it is desirable that the conclusions of a study remain the same even if the circumstances change, in order to increase the chance of a successful translation of the findings from preclinical to clinical studies, as the end goal of pharmaceutical research is to provide a new treatment. Thus, in the long-term, the focus on improving and quantifying reproducibility can also positively impact translating research from preclinical to clinical studies and, consequently, improve the efficiency of the drug development process.

3.6 Statistical reproducibility

Up to this point, this chapter has categorised definitions of reproducibility, presented reasons for low reproducibility and suggestions on how to improve reproducibility, and discussed reproducibility within the context of preclinical research. This section considers reproducibility from the perspective of statistics. This section aims to provide a concise summary of the debates on the topic of statistical reproducibility addressed in the literature. For example, statistical reproducibility has been discussed in depth in 'Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop' [152]. These proceedings of the workshop considered important questions in the research area but it did not give a summary of the existing metrics that are aimed at validating reproducibility and quantifying statistical reproducibility. This task will be pursued in Sections 3.7, 3.8 and 3.9. Section 3.6.1 discusses different definitions of statistical reproducibility and questions that have been raised regarding statistical reproducibility and Section 3.6.2 considers the debate on whether or not to use p-values.

3.6.1 What is statistical reproducibility?

Similar to the term *reproducibility*, the term *statistical reproducibility*, reproducibility probability or replication probability, is not clearly defined. The first insights related to statistical reproducibility were provided by Goodman [91], who highlighted a misconception regarding the p-value. Goodman [91] questioned the claim that a small p-value improves the credibility of the test result and argued that the replication probability may be smaller than expected. Although Goodman used the term replication probability rather than reproducibility probability, his definition is similar to the definition of reproducibility adopted in this thesis. Goodman [91] defined it as the probability of observing another statistically significant result in the same direction as the first one, if an experiment was repeated under identical conditions and with the same sample size, which is consistent with Reproducibility Type C. Senn [182] agreed with Goodman that the p-value and replication probability are different measures and that inconsistency between test results from individual studies may be expected. However, he disagreed with Goodman's claim that the p-value may overstate the evidence against the null hypothesis [91], both under the Frequentist and the Bayesian framework. According to Senn [182], under the Frequentist framework, p-value is the most rigorous possible type I error rate that could be considered and still lead to the rejection of the null hypothesis. Under the Bayesian framework, it could be argued that the p-value corresponds to a particular Bayesian posterior probabilities. Nevertheless, Senn [182] recognised that a link between the p-values and replication probability should be recognised. This thesis uses the term reproducibility probability (RP) instead of replication probability and it will return to this topic in Chapter 4.

Miller [149] argued that there are two interpretations of the replication probability and that in both cases the probability is unknown. Miller called them the aggregate and the individual replication probability [149, p.618]. According to Miller, the former term refers to experiments being performed by different teams of researchers with varying conditions, which corresponds to Reproducibility Type E, whereas the latter term refers to experiments being carried out by a particular individual under exactly the same conditions, which corresponds to Reproducibility Type C and to Goodman's definition of statistical reproducibility. Miller discouraged researchers from attempting to estimate both types of replication probabilities, as, according to him, the initial data provide very little informa-

tion about the RP in the follow-up experiment [149, p.629]. This is something we disagree with; a statistician uses data for inference, hence, it contains further information.

Stodden [196] had a different approach to the use of the term *statistical reproducibility*. She described it as conception about how statistics affect the likelihood of a scientific result being reproducible and how they contribute to the study and the quantification of reproducibility [152, p.4]. Stodden also used this term to refer to the situation when flawed statistical analysis or experimental design leads to the failure to replicate the experiment [196]. The positive side of this definition is that it emphasises the importance of appropriate use of statistics in experiments. However, this definition generalises *statistical reproducibility* to any discussion regarding statistics and reproducibility, and it cannot be classified as any of the Reproducibility Types introduced in Section 3.2.

The debate on statistical reproducibility raises a variety of questions. In the proceedings of the workshop on 'Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results' by National Academies of Sciences, Engineering, and Medicine (NASEM) [152], one of the questions focused on what study designs and appropriate metrics can be used to quantify reproducibility of scientific findings. The proceedings of NASEM [152] mainly concentrated on the variability across studies, on how to assess this variability and on what degree of variability leads to worries about the lack of reproducibility. Indisputably, variability is an important factor in the statistical reproducibility debate. Lomax [152, p.15] explained that it is important to recognise which aspects of variation can and which cannot be controlled.

Exchangeability of random variables forms part of the variability discussion. De Finetti's Theorem [107] states that exchangeable observations are conditionally independent. It means that variables can be swapped around in the sequence, and following this their joint distribution does not change. Exchangeability can never be verified, but statisticians still make the assumption of exchangeability under the guidance of practitioners. In the reproducibility debate, it is important to ask whether or not one can assume exchangeability. This thesis proposes that exchangeability could be assumed when the replicate experiment is carried out under the same conditions. This work assumes exchangeability in the nonparametric predictive inference (NPI) framework, which is used to quantify NPI reproducibility in Chapter 4. Thus, exchangeability can only be assumed

for Reproducibility Type C. It is arguable whether exchangeability, or some extent of exchangeability, can also be assumed for Reproducibility Type D. Exchangeability can no longer be assumed for Reproducibility Type E, where the experiments are carried out under different conditions. These scenarios are outside the scope of this thesis, as this work focuses on the quantification of Reproducibility Type C. Nevertheless, a question of how to address variability in scenarios where exchangeability cannot be assumed is of further research interest.

The proceedings of NASEM also discussed how statistics, in particular the choice of study design and analysis, can affect reproducibility of scientific results, and how reproducibility can be enhanced via structural and analytical approaches [152, p.3]. These questions address statistical causes of poor reproducibility and suggestions for improvements, both of these have been addressed in Section 3.4.

Lastly, there is an important question: Within what framework should statistical reproducibility be assessed? BinHimd and Coolen [31, 53] considered reproducibility as a predictive problem and provided a frequentist approach, nonparametric predictive inference, to solve it. This thesis adapts their approach to statistical reproducibility. Within the Bayesian framework, predictive inference has been discussed by Billheimer [30]. With a view to improve reproducibility, Billheimer [30] proposed predictive inference to predict observables. According to Billheimer [30], statistical modelling should predict observable quantities and events, based on the current data and other applicable information, rather than form inferential conclusions through hypothesis tests or estimation of parameters. Billheimer promoted that instead of focusing on unobservable parameters, attention should be centred on observable events. This view is in alignment with the approach to statistical reproducibility presented in this thesis, however, this work suggests using NPI framework instead of Bayesian framework, as NPI does not make many assumptions about the data whereas Bayesian framework does.

3.6.2 The p-value and the statistical significance

Concerns regarding reproducibility of research results is interlinked with the ongoing debate about whether or not to use p-values [30]. In hypothesis testing, which is a method of statistical inference, p-values are used to make dichotomous decision about whether

the data support a particular hypothesis. Depending on the p-value, test outcomes are labelled statistically significant or non-significant. The most commonly used threshold value for the p-value is 0.05. The p-value is the probability of obtaining the same or a more extreme value for the test statistic, under the assumption that the null hypothesis is correct. The American Statistician (TAS) [206] suggested abandoning the concept of *statistical significance* in scientific research. The grounds for this suggestions are that the concept of statistical significance is misinterpreted by many, that it can cause erroneous beliefs and poor decision making, and that it stops statistically insignificant results from being published. Furthermore, statistical significance does not imply truth, yet many researchers and bodies equate it with truth [206]. The editorial [206] stated that it is not enough to have directions, such as "Don't believe that an association or effect exists just because it was statistically significant", but that the p-values should not be dichotomised, i.e. test outcomes should not be labelled as statistically significant or non-significant, and the word statistically significant should not be used. TAS [206] suggested that rather than stating the p-value, its meaning should be described in words.

Fisher introduced p-values for the use at the exploratory stage to see if the experiment findings should be further investigated [156, 206]. They were not meant to lead to a dichotomous decision making rule, reject or not reject the null hypothesis. The dichotomous nature of significance testing often leads to p-hacking, the misreporting of true effect sizes by researchers who want to publish and need significant results, as discussed in Section 3.4. Similarly to TAS, Amrhein et al. [6] argued that dependence on statistical significance threshold can be misleading, and they suggested not using statistical significance thresholds and reporting only precise p-values.

Amrhein et al. [6] argued that conclusions should not be based solely on whether the p-values are significant or non-significant. Other metrics, such as the effect size and power, are equally important in the statistical analysis of tests. Amrhein et al. [6] also addressed the problem of making over-confident claims based exclusively on p-value. Nuzzo [156] also highlighted that effect sizes are often ignored and the research focus is on whether there is an effect rather than on how big the effect is, while the latter question is often more important. Nuzzo [156] discussed the problem of overrelying on p-values in decision making. Halsey et al. [99] discouraged analysis based mostly on p-values because of

"the wide sample-to-sample variability in the p-value" [99]. They proposed that the dichotomous yes-or-no decision should be reached using a variety of measures, in particular the effect size estimates and their 95% confidence intervals [99]. Colquhoun [46] raised the problem of high false discovery rate for p-value around 0.05 in significance testing. He illustrated this on tree diagrams for simple testing procedures and he explores it further via simulations; the false discovery rate is the ratio of the number of false positive results to the total number of positive test results. The false discovery rate is also high for tests with low statistical power. Chapter 4 will show that statistical reproducibility is low for test decisions with a p-value close to the threshold. There is a possible link between this and Colquhoun work.

Halsey [98] offered four alternative analysis approaches to augment or replace the p-value. First, he discussed the augmented p-value - augmented with information about its variability. He suggested the p-value prediction interval as a possible tool to do so. The prediction interval characterises the uncertainty of the p-value of a future replicate study [60]. It is questionable whether augmented p-values will not create more confusion as their interpretation is not straightforward and their calculation is based on p-values. Secondly, Halsey suggested estimating effect sizes and their confidence intervals [98]. Thirdly, Halsey suggested the use of Bayes factors instead of p-values as more intuitive metrics for interpretation. Fourthly, Halsey suggested using the Akaike information criterion for model assessment, which is outside the scope of this chapter's discussion. Being aware and using alternative methods for statistical analysis gives decision-makers more flexibility and more tools to make decisions. However, these tools do not replace p-values as they are different measures and communicate different messages.

Macnaughton [140] disagreed with the claims made by TAS [206], in particular, that abandoning statistical significance will lead to fewer false-positive errors in scientific research, and that it will enable easier replication of scientific research results [140]. According to Macnaughton [140], science and statistics aim at separating signal from noise in data and the p-value is a useful tool for determining whether the studied effect exists in the population [140]. Unfortunately, false-positives still persist in published research, i.e. a p-value which implies that there is evidence for the alternative hypothesis, but in fact the null hypothesis is true. Macnaughton [140] argued that the critical threshold

value provides a balance between the rates of false-negative errors, false-positive errors, and costs. Macnaughton acknowledged that some people may manipulate p-values (either because of a lack of knowledge or on purpose so that they can publish) and this is harmful to science. Macnaughton also pointed out that if researchers obtain a p-value above the critical value of a relevant journal and if they believe that the studied effect exists and it is important, then the researchers should create a more powerful research design and repeat the study to see if they can get convincing evidence for the existence of the effect. Ioannidis [120] also argued that significance is essential for activity in both science and non-science and that some filtering process is helpful to avoid drowning in noise.

Benjamin et al. [25] proposed to change the default p-value threshold for statistical significance from 0.05 to 0.005 for new discovery claims, to improve reproducibility and to label novel findings with p-values between 0.005 and 0.05 as suggestive evidence. Reproducibility was not explicitly defined by Benjamin et al. [25], it could be assumed that they referred to Reproducibility Types C, D or E, or a combination of these. Benjamin et al. [25] did not propose that this new threshold is used for decisions on whether to publish or not. Similarly, the proceedings of the workshop by NASEM [152, p.48–55] discussed the benefits of increasing the threshold for demonstrating statistical significance, through p-values or Bayes factors. It is doubtful whether this would increase reproducibility because p-values and reproducibility probability are different measures and there is inconsistency between test results from different studies, as has been discussed by Senn [182] and Goodman [91].

Leek and Peng [136] identified that there are more important discussions than the question of whether or not to use p-values. It is more important to focus on the improvement of researchers' education in statistics and evidence-based data analysis, teaching them to use statistical analysis correctly. We agree with their point of view; the p-value forms only a small part of the experiment, which follows experimental design, collection and handling of data, and summary statistics, and the problem of a lack of reproducibility in science cannot be solely blamed on the p-value.

3.7 Replicate studies

Following all the documentation of the study, carefully checking the study design, code, data analysis and other aspects of the study, making sure that there is no error in alignment with Reproducibility Type A. Reproducibility Type B necessitates the use of the data of the original study but to apply a different analytical method and see whether the same conclusions were reached through this process.

As discussed in Section 3.2, for Reproducibility Types C, D and E, there are two scenarios: First, both the original and the replicate experiments have been performed and a need to assess whether the conclusions of the original study are reproduced in the replicate study arises. Secondly, only the original experiment has been performed and the reproducibility of the experiment is assessed, based on the data and the statistical analysis. The first scenario has received more attention in the literature and several approaches have been developed for assessing whether replication of an experiment has been successful. This section gives an outline of some of those methods. Different metrics are used across different fields and for different data. There is no agreement on the use of one metric. This section will outline some approaches from the literature aimed at assessing reproducibility via validating outcomes of replicate studies. The second approach, in particular in combination with Reproducibility Type C, is the main focus of this thesis. Some available metrics for the quantification of reproducibility from the perspective of the second approach will be discussed in Section 3.8, following presentation of NPI reproducibility in Section 3.9.

3.7.1 Reproducibility Projects: Psychology and Cancer biology

As has been stated in the beginning of this chapter, Errington et al. [81,82] carried out the Reproducibility Project: Cancer Biology, in which the replicate studies were conducted for 50 experiments from 23 original papers. Open Science Collaboration [157] replicated 100 experimental and correlational studies published in three psychological journals. This section discusses the assessment of reproducibility in both Reproducibility Projects.

Errington et al. [82] described seven methods for the assessment of replication:
(i) statistical significance: whether the p-value is less than 0.05 for the original positive

results or whether the p-value is greater than 0.05 for the original null results; (ii) original effect size in the replication 95% confidence interval; (iii) replication effect size in the original 95% confidence interval; (iv) replication effect size in the original 95% prediction interval; (v) meta-analysis combining original and replication effect sizes, leading to p-value less than 0.05 for the original positive results or to p-value greater than 0.05 for the original null results; (vi) comparing whether the results had the same direction - in the evaluation of representative images the original and replicate outcome can have the same direction but a different statistical significance; (vii) comparing whether the replication effect size is less than or equal to the original effect size. A replicated study was assessed as successful if majority of the criteria (i) - (v) were satisfied (3 or more out of 5). The other two criteria, (vi) and (vii), were not included in this assessment of a successful replication, as they do not work for null effects, i.e. cases when the null hypothesis is not rejected. The comparison of effect sizes showed that the median of effect sizes in the replication studies was 85% smaller than the median of effect sizes in the original experiments, and 92% of the replication effect sizes were smaller than the original effect sizes. Moreover, the original null effects were replicated for 80% of the original tests, whereas the positive findings were replicated for only 40% of the original tests.

Open Science Collaboration [157] evaluated reproducibility via the following criteria: significance and the same p-value cut-off point, effect sizes, subjective assessment of replication teams, and meta-analyses of the effect sizes. They concluded that while 97% of the original studies had a p-value below 0.05, only 36% of the replication studies had a p-value below 0.05.

Patil et al. [158] highlighted the problem that the p-value cut-off points do not account for variation [158]. Patil et al. [158] instead suggested the consideration of the effect expected in the replication study, examining the original effect. Patil et al. [158] defined the 95% prediction interval, which can be calculated via Equation (3.1).

$$\hat{r}_{\text{original}} \pm z_{0.975}\sqrt{\frac{1}{n_{orig}-3}+\frac{1}{n_{rep}-3}} \tag{3.1}$$

where $\hat{r}_{\text{original}}$ is the correlation estimate in the original study, n_{orig} and n_{rep} are the sample sizes in the original and the replication study, respectively; and $z_{0.975}$ is the 97.5% quantile of the Normal distribution [158].

Patil et al. [158] warned that a small sample size leads to a wide prediction interval; and thus the assessment about the replication study could be non-informative for small sample sizes. Patil et al. [158] pointed out that in the Reproducibility Project: Psychology [157] by Open Science Collaboration, the replication study effect sizes were smaller than the original study effect sizes due to publication bias. This observation is in line with the observation made in Reproducibility Project: Cancer Biology [82].

3.7.2 High throughput experimentation

In high throughput experimentation (HTE) automated equipment is used to run a large number of tests simultaneously. Parallelisation is the key principle of HTE. High throughput experimentation is, for example, used in biological science laboratories to rapidly screen millions of samples. Assessment of reproducibility is a highly discussed topic in high throughput experimentation, where the replicate study often has a different sample size to the original study. In the replicate studies, only signals, that were positive, interesting or significant in the original study, are studied, thus, the sample size and design in the replicate study differs from the original study. Moreover, scientists sometimes introduce test compounds in the replicate study that have similar characteristics to those selected as significant in the primary screen.

The metrics used to quantify reproducibility in HTE are the r-value [102], irreproducible discovery rate (IDR) [139] and maximum rank reproducibility (MaRR) [165]. The r-value is briefly described, a detailed discussion of these metrics is outside the scope of this thesis. This short survey of available metrics aims to illustrate that this type of assessment has received considerable attention in the literature. Both Li et al. [139] and Philtron et al. [165] named Spearman's pairwise rank correlation as a commonly used method for assessing reproducibility in HTE, however, both sources agreed that it is not the most suitable method as Spearman's pairwise rank correlation's properties depend on how stringent the requirements for inclusion of genes are.

In the field of genomics, assessing whether findings from a primary study are replicated in a follow-up study has been explored [35, 103]. The terminology used is *replicability*, findings being replicated in another study. The studies conduct large-scale searches for rare true positives; one study is simultaneously examining many features. In the context

of genome-wide association studies, the follow-up studies often examine only features that were identified as significant in the primary study.

For the test scenarios described above, Heller et al. [102] introduced the r-value as a metric to quantify the strength of replication [102], i.e. evidence against findings from a primary study being replicated in a follow-up study. A smaller r-value means stronger evidence in favour of replicability [187]. The Benjamini-Hochberg procedure can be used on the reported r-values to control the false-discovery rate (FDR). Heller et al. [102] defined the FDR r-value for feature i as the lowest FDR level at which the finding is among the replicated ones. Heller, Bogomolov and Benjamini offered an online calculator of the r-value at `http://www.math.tau.ac.il/~ruheller/App.html`. Meta-analysis is often used in genome-wide association studies, however, Heller et al. [102] argued that meta-analysis, pooling results across studies, is not an assessment of replicability and they suggested to add the r-value to the statistical analysis. Further discussion of the r-value is outside the scope of this thesis.

3.7.3 Agreement indices

Assessment of whether a replicate study reached the same conclusions as the original study, in accordance to Reproducibility Type C and Type D, has also been assessed via agreement indices. Barnhart et al. [21] compared various agreement indices: the Pearson correlation coefficient, the mean-squared deviation, the intraclass correlation coefficient, kappa statistic, the concordance correlation, the within-subject coefficient of variation, coefficient of individual agreement, limits of agreement, coverage probability, and total deviation index. They identified the coverage probability as the preferred index for assessing agreement, because it can be applied to both continuous and categorical data, it is intuitive and easy to compute. These metrics are not described separately as they are not relevant to the rest of this thesis.

3.7.4 Reproducibility from a Bayesian perspective

Reproducibility has been assessed from a Bayesian perspective [22,101,188]. For example, Held [101] introduced the sceptical p-value (p_S), a quantitative measure for *replication success*. The term *replication success* is not explicitly defined by Held [101], we assume that it means that the findings of the original experiment are validated in the replicate experiment. The technique is suitable for tests which employ frequentist analysis. It considers p-values, sample and effect sizes of both the original and replication study. The method determines the largest confidence level $1 - p_S$ for the original confidence interval, at which replication success can be declared at level p_S [101]. The author preferred this method to meta-analysis because, according to him, exchangeability assumptions are not appropriate [101]. Held's argument is that, via the conduct of a replication study, researchers challenge the findings of the original study, which is an asymmetric task. The problem we encounter with this method is that the term *replication success* is not clearly defined and the definition of the sceptical p-value involves this term.

3.8 The quantification of statistical reproducibility

Section 3.7 discussed metrics assessing reproducibility in situations where both the original and the replicate experiments have been carried out. This section focuses on metrics which are calculated after only the original study has been carried out. These metrics relate to the probability of getting the same decision in a follow-up study. This view of reproducibility is in alignment with Goodman's [91] definition of statistical reproducibility and Billheimer's [30] approach to predictive analysis. In the literature, less attention is paid to this approach to statistical reproducibility. This thesis focuses on such assessment.

3.8.1 Confusing reproducibility with other statistics

We have noticed that some researchers interpret p-values, effect sizes or confidence intervals as measures of reproducibility. For example, Soderberg [152, p.58] stated that p-values or effect sizes are examples of different ways of measuring reproducibility. Boos [152, p.49] matched p-values with reproducibility probability, e.g. $p = 0.01$ equals to $\widehat{RP} \approx 0.73$ and $p = 0.0001$ equals to $\widehat{RP} \approx 0.97$. It is unclear how these $\widehat{RP}$ values are defined or

calculated. Cumming [60] argued that confidence intervals contain information about replication. We disagree that p-values, effect sizes or confidence intervals are measures of reproducibility as they have a clear definition in statistics, and reproducibility or related terms are not part of those definitions; these are different concepts.

3.8.2 Peculiar metrics

In the literature, there are peculiar measures of reproducibility, such as Posavac's t_{rep} and Kileen's p_{rep}. Both metrics are linked to significance testing. According to Posavac [169], the probability of *statistically significant exact replication*, t_{rep}, can be calculated by subtracting the minimum difference for a statistically significant t-statistic from the difference in means observed in the initial study. Posavac presented a graphical method for calculating the probability of an exact replication being less than 0.05 for a two-tailed test [169]. However, it is not clear from the article how this would quantify the probability of the next experiment yielding the same conclusion. Because of the vagueness of the approach, it is unclear how to apply it in practice.

Killeen [132] argued that the probability of replicating an experiment can be estimated using the statistic p_{rep}. He defined p_{rep} as the replicate effect which is of the same sign as the effect found in an original experiment [132]. Killeen was motivated by the fact that the p-value is commonly misinterpreted. According to Killeen [135], p_{rep} can be estimated by viewing it as a function of the p-value (denoted by p), using the following formula:

$$p_{rep} \approx [1 + (\frac{p}{1-p})^{2/3}]^{-1} \tag{3.2}$$

Maraun and Gabriel [143] pointed out that Killeen's calculation and interpretation of p_{rep} and of the concept of reproducibility probability contain errors. Nevertheless, they credit Killeen's claim that replicability should play a key role in the assessment of empirical results [143]. Lecoutre et al. [135] also recognised that p_{rep} is incorrectly defined, because of the confusion between 1-tailed and 2-tailed p-values. Another problem with Posavac's and Kileen's calculations of reproducibility is that both of these metrics are dependent on p-values, which are not measures of reproducibility, as explained in Section 3.8.1.

3.8.3 Estimated power approach

De Capitani and De Martini [66–68] adopted Goodman's definition of reproducibility probability, i.e. the probability of obtaining the same test result in a second, identical experiment. This corresponds to Reproducibility Type C, but they considered it as an estimation problem instead of a prediction problem.

De Capitani and De Martini [66–68] equated reproducibility probability to the true power of a statistical test. Their method is called the *estimated power approach* [184] and has been presented for the t-test, Wilcoxon rank-sum test [66] and they also developed reproducibility probability estimation for other nonparametric tests [67]. Shao and Chow [184] also advocated the estimated power approach. De Capitani and De Martini [67] argued that their methods provides useful information for evaluation of the stability of statistical test results. It is unclear what is the precise definition of the stability of test results and what is the benefit of the estimated power approach.

De Capitani and De Martini argued that many clinical trials cannot be done more than once or twice, mainly because of their budgets and time constraints [65, p.1]. However, for an experiment to be scientifically valid, it is often required that it is reproducible. De Capitani [65] argued that in such cases reproducibility of the experimental conclusions should be addressed as reproducibility of statistical significance [65] and this should be evaluated using reproducibility probability. We disagree with their statement as we believe the interest should be in reproducibility of conclusions rather than reproducibility of statistical significance and the two cannot be equated.

Furthermore, De Capitani and De Martini only considered reproducibility in the case of the null hypothesis being rejected, while this thesis provides predictive inference for reproducibility for both cases: when null hypothesis is rejected and when it is not rejected.

3.8.4 G $\times$ L adjusted p-value

It is hard to achieve standardisation in preclinical research and there has been a shift to embracing variability, as discussed in Section 3.5. In line with Reproducibility Type E, all conditions cannot be the same in the replicate experiment. Kafkafi et al. [128] described genotype-by-laboratory interaction (G $\times$ L) adjusted p-value, a metric that is aimed at

accounting for variability in genotype influenced by environment. G × L adjusted p-value indicates the probability of replicating the result in additional laboratories [128].

The sensitivity of strains of mice - animals with identical genetics - to the environment is assessed by collecting results about different strains from different laboratories and determining how consistent is the phenotype, i.e. the set of observable characteristics. The G × L adjusted p-value is derived by estimating the interaction noise $\sigma^2_{\mathrm{G\times L}}$ from studies of a number of strains of mice in different laboratories. This provides information on the extent to which the p-value needs to be adjusted. For example, if the strain is very susceptible to the environment, the p-value adjustment is greater. The International Mouse Phenotyping Consortium (IMPC) strived to promote a public database of mutant lines of mice that could be available to all laboratories. The random lab model (RLM) adds the interaction noise $\sigma_{\mathrm{G\times L}}$ to the animal noise to create a base for determining phenotype differences [128]. The power is subsequently lowered and confidence interval of the estimated effect size is widened, accordingly, to ensure replicability. In theory, scientists could calculate G × L-adjusted p-values and confidence intervals. However, the method does not appear to be developed for a wider-use application. The authors [128] of the article claim that reporting G × L-adjusted p-values and confidence intervals alongside the usual p-values and confidence intervals would increase replicability in preclinical research but they do not present reasons.

This approach seems, at first sight, appealing, as the calculation of the G × L adjusted p-value takes into account results from a variety of laboratories. However, even the G × L adjusted p-value is susceptible to errors. The feasibility of this method is related to the question whether it is possible to accurately estimate the G × L variability and if it is reasonable to trust this estimate. The variability in animal testing is complex, it does not only depend on the mouse batch and a particular laboratory, but also on the person who runs the experiment, the time of the day or the year, and the environment conditions, as discussed in Section 3.5. However, this question is outside the scope of this thesis, as it is not a statistical question.

3.9 NPI reproducibility in the context of the literature

Statistical reproducibility has received attention in the literature. The main focus has been on the variance between the original and the replicate experiments, discussed in Section 3.7, and the debate on the p-values, discussed in Section 3.6.2. Most available metrics validate conclusions in a replicate experiment in scenarios where both the original and the replicate experiments have been performed, and the question of interest is whether the conclusions of the original experiment have been reproduced in the replicate experiment, and some of these metrics were discussed in Section 3.7.

However, the quantification of statistical reproducibility in the case where only the original test scenario has been carried out, has received notably less attention. Metrics relating to this scenario were discussed in Section 3.8. However, none of those metrics focused on data. They either paid attention to the metrics of the current data analysis or to the variability caused by environmental factors. The data aspect is important because data of the original experiment can reveal information about the variability.

This thesis interprets statistical reproducibility as a prediction problem. The data driven, predictive approach shows resemblance to Billheimer's approach, although NPI is a frequentist framework, while Billheimer was advocating a Bayesian approach. Reproducibility probability is defined as the probability of the event that, if a test was repeated under identical circumstances and with the same sample size, the same test outcome would be reached, which resembles Goodman's definition of statistical reproducibility and reflects Reproducibility Type C. NPI reproducibility has been introduced in Section 1.5. As noted in Section 1.4, Hill's assumption, on which the NPI is based, requires that random quantities are exchangeable. Although exchangeability assumption may not be realistic in practical scenarios, it is possible because under the NPI analysis, as the actual experiments are not redone in practice.

Section 3.4, inter alia, considered what does the term low reproducibility mean. In line with this thesis approach to statistical reproducibility, low reproducibility can refer to a low probability of the event that, if a test was repeated under identical circumstances and with the same sample size, the same test outcome would be reached [52]. This interpre-

tation of low reproducibility has not received much attention in the literature. However, it can be quantified under the nonparametric predictive inference (NPI) framework, and this thesis will concentrate on this interpretation of reproducibility in Chapter 4.

3.10 Concluding remarks

This chapter introduced the main debates in the area of reproducibility. There is no universally accepted definition of the term *reproducibility*. Various definitions of reproducibility and related terms available in the existing literature were discussed and classified into five types of reproducibility in Section 3.2. It was shown that sometimes different definitions are used for the same term and sometimes the same definition is used for different terms; some definitions are not clear; and often the term *reproducibility* is used without being explicitly defined. Reasons for low reproducibility and suggestions for improving reproducibility offered in the literature were outlined in Section 3.4. A lot of the solutions simply entail adhering to good scientific practice and using appropriate statistical, experimental and documentation methods, and collaboration of different stakeholders. This chapter briefly discussed reproducibility in relation to preclinical research in Section 3.5, presenting ethical issues and other reproducibility challenges in preclinical research, listing possible solutions to low reproducibility in preclinical research, offered in the literature, focusing on the shift from striving for homogeneity to embracing variability.

Then, Section 3.6 focused on the debates relating to the statistical reproducibility. Similarly to the concept of reproducibility, statistical reproducibility is not a clearly defined term. Goodman [91] defined reproducibility as the probability of observing another statistically significant result in the same direction as the first one, if an experiment was repeated under identical conditions and with the same sample size. This definition of statistical reproducibility is adopted in this thesis.

Statistical discussion of reproducibility has focused on the variability across studies and ways of controlling this variability. This section raised further important questions, such as whether the assumption of exchangeability is important for the quantification of reproducibility and within what framework should reproducibility be assessed. Linked to the reproducibility debate has been the ongoing discourse on whether to use p-values.

This chapter presented arguments for and against the use of p-values and concluded that although there are many issues and problems associated with p-values, there is no clear and straightforward alternative to p-values that could be widely adopted by researchers.

Section 3.7 outlined some of the metrics determining whether reproducibility (often called replication) has been successful in scenarios where both the original and the replicate experiments have been carried out. This thesis focuses on quantifying statistical reproducibility in the case where only the original experiment had been carried out. This topic has received less attention in the literature. Section 3.8 presented a summary of metrics for quantifying reproducibility that are available in the literature. Estimated power relates to current statistics. G $\times$ L adjusted p-value also falls into this category, however, it tries to incorporate into the p-value the variability caused by the interaction with environment; it assumes the conditions cannot be the same in a replicate experiment.

Section 3.9 identified a gap in the current debate, i.e. the consideration about what can data from the original study reveal about statistical reproducibility. This chapter proposed addressing reproducibility as a predictive problem and using NPI framework, introduced in Section 1.5, to quantify it. Chapter 4 will further develop on NPI reproducibility, presenting practical implementation of NPI reproducibility for the Wilcoxon Mann-Whitney test and the t-test.

NPI can quantify statistical reproducibility, which opens up additional questions. An important question is what should a decision-maker do when reproducibility is low? A statistician would most likely advise that in such cases an experiment should be re-run, preferably with larger sample sizes. However, as shown in Section 3.5, there are often ethical and financial constraints that make the replication of the experiment hard. It is of future research interest to present an action plan for cases where NPI reproducibility is low.

In line with Reproducibility Type C, it is assumed that the replicate experiment would be carried out under the same conditions. However, if the sample size of the future sample increased or decreased, and everything else stayed the same, could exchangeability still be assumed and how would this affect NPI reproducibility? In theory, NPI framework allows to make predictions based on the assumption of future sample being of a different sample sizes than the original sample, however, the approach to exchangeability in such

scenario is a topic for future research.

Moreover, the method for quantifying reproducibility would allow a practitioner to calculate NPI reproducibility of an experiment, then carry out a replicate of that experiment and then calculate NPI reproducibility for this replicate experiment. This leads to the question: If the experiment was repeated, what conclusions could be made about reproducibility of a second repeat of the experiment, based on the calculated NPI reproducibility for the original and the replicate experiment? Application of this in practice would present more directions for further research action.

Chapter 4

Statistical reproducibility for pairwise tests in preclinical research

4.1 Introduction

Sections 1.5 and 3.9 introduced NPI reproducibility. NPI reproducibility probability is the probability of the event that, if a test was repeated under identical circumstances and with the same sample size, the same test outcome would be reached. This chapter presents advances on the topic of statistical test reproducibility with relevance to the t-test and the Wilcoxon Mann-Whitney test (WMT), two statistical tests commonly used in preclinical research. As explained in Sections 1.4 and 1.6, NPI analysis traditionally involves calculating lower and upper probabilities by considering all the orderings. BinHimd [31] encountered a problem when calculating precise lower and upper reproducibility probabilities for the WMT: it is computationally hard to derive such lower and upper probabilities for practical data sets since the number of orderings to consider grows exponentially as the number of the original data points increases. She dealt with the problem by employing NPI bootstrap, presented in Section 2.3.3, to calculate approximate NPI bootstrapped reproducibility probability (NPI-B-RP). This thesis develops further on BinHimd's work. This work calculates NPI reproducibility for a parametric test, the t-test, which introduces further complications: computing the minimum and maximum values of the t-test statistic for m future observations with given ordering O_i is difficult, because this statistic depends both on the sample mean and variance. Therefore, in this thesis, estimates for

reproducibility probabilities are calculated instead of the precise lower and upper reproducibility probabilities.

Two implementations of NPI for reproducibility are available for estimating reproducibility probability for the WMT and the t-test: NPI bootstrap, presented in Section 2.3.3, and the sampling of orderings method, introduced in Section 1.6. NPI bootstrap provides a point estimate whereas the sampling of orderings method calculates estimates of lower and upper reproducibility probabilities.

For both approaches, NPI reproducibility is applied to a real-life scenario of a preclinical experiment, which involves multiple pairwise comparisons of test groups, where different groups are given a different concentration of a drug. The aim of the experiment is to decide the concentration of the drug which is most effective. This test scenario is introduced in Section 4.2.

Implementation of NPI bootstrap is more straightforward and the focus of this chapter is on this implementation. Section 4.3 introduces the NPI bootstrap implementation for the t-test. NPI-bootstrap is employed to quantify a bootstrapped estimate for the statistical reproducibility of the pairwise t-test. To explore NPI reproducibility for the t-test, simulations both under the null and alternative hypotheses are carried out and then reproducibility for the test scenario is calculated. In both simulations and the application scenario, the relationships between reproducibility and two test statistics, the Cohen's d and the p-value, are studied. Reproducibility of the t-test is also compared with reproducibility of the WMT.

Then, Section 4.4 examines reproducibility for the final decision of choosing a particular dose in the multiple pairwise comparisons scenario. This topic has not yet received much attention in the literature but provides interesting insights regarding statistical reproducibility. This chapter will show that statistical reproducibility for the final decision is notably lower than reproducibility for separate pairwise comparisons.

The sampling of orderings for the likelihood test was presented by Marques et al. [55,144,145]. Estimates of lower and upper reproducibility probabilities can be calculated via sampling of a particular number of these orderings and carrying out the reproducibility analysis on these orderings. Section 4.5 presents the methodology for the WMT. It is more challenging to do so for the t-test than for the WMT. Nevertheless, Section 4.6 presents

heuristics for calculating NPI-RP estimates for the t-test.

Section 4.7 explores NPI reproducibility for the rate of growth measure, a metric commonly used in preclinical research. This measure is calculated from measurements at various time points, rather than just at the end point. Three different studies are presented. In these studies the data sets cannot be assumed to come from a Normal distribution, thus, the WMT is applied, alongside the growth rate (GR) inhibition significance analysis.

The chapter concludes with a summary of the findings and with the formulation of future research questions in Section 4.8.

4.2 Motivating preclinical test scenario

This section introduces the motivation test scenario, which uses real data. The experiment assesses 6 concentrations of a drug; A is the control group and B-F are groups given increasing concentrations of the drug. For each group, there is one measurement available for each individual. The measurement is such that the lower the recorded value is, the better the drug performs at that concentration. The data have been log transformed to meet the t-test assumption of Normality; they are presented in Table 4.1 and Figure 4.1.

Dose						
A	B	C	D	E	F	D'
0.7450	0.5148	0.1088	0.0133	-0.1221	-0.1946	0.4033
0.7513	0.5280	0.1732	0.0265	-0.1010	-0.0520	0.4087
0.8484	0.5546	0.1896	0.0302	-0.0519	-0.0417	0.4103
0.8584	0.5553	0.2202	0.0444	-0.0436	-0.0039	0.4163
0.8728	0.6265	0.2352	0.0882	-0.0200	0.0076	0.4354
0.8964	0.6315	0.2697	0.1461	-0.0182	0.0196	0.4624
0.9053	0.6890	0.3298	0.1545	-0.0104	0.0512	0.4665
1.0981	0.7605	0.4150	0.1585	0.0879	0.1540	0.4684
	0.7843	0.4234	0.2638	0.1390	0.2247	0.5232
	0.8173	0.4401		0.1945		

Table 4.1: Log transformed data for each dose (D' replaces D in Section 4.4.2)

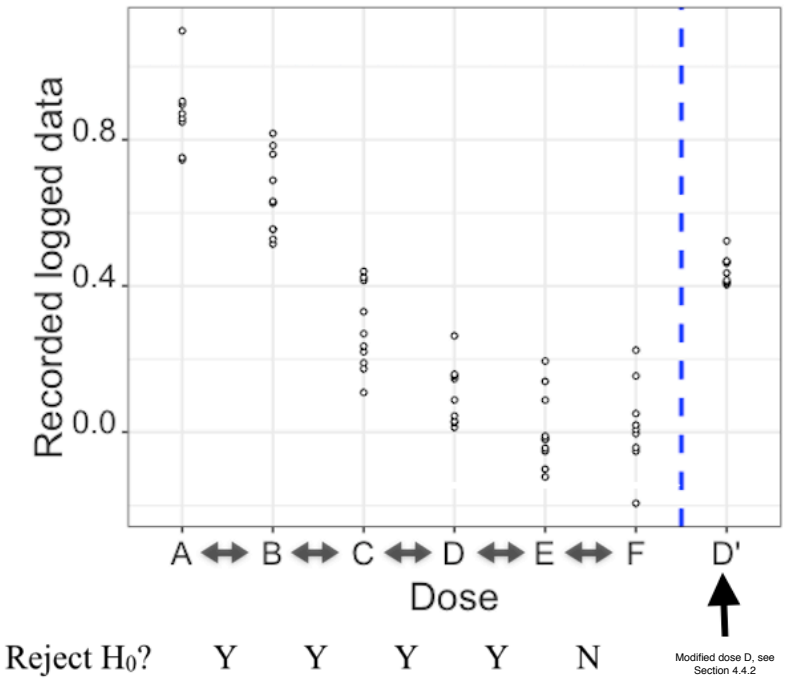


Figure 4.1: Log transformed data for each dose and outcomes of pairwise comparisons (D' only used in Section 4.4.2)

Five pairwise comparisons are carried out between adjacent concentrations of the drug (A vs. B, B vs. C, C vs. D, D vs. E, E vs. F). For each pairwise comparison, the question of interest is if the dose with a higher concentration is performing better than the dose with a lower concentration. In each pairwise comparison, the upper-sided t-test with the equal variance assumption is applied. Let μ_H denote the population mean for the dose with the higher concentration and μ_L the population mean for the dose with the lower concentration. The null hypothesis is $H_0 : \mu_L = \mu_H$ and the alternative hypothesis is $H_1 : \mu_L > \mu_H$. The significance level α is equal to 0.05. For each pairwise comparison, the test outcome is either to reject (Y) or to not reject (N) the null hypothesis.

The results of multiple pairwise comparisons for the data presented in Table 4.1 are YYYYN, indicating that the null hypotheses are rejected for all pairwise comparisons except for last one, E vs. F. As seen in Figure 4.1, as the dose increases, the measurements tend to decrease until dose E.

Note that the nonparametric counterpart to t-test, the Wilcoxon Mann-Whitney test

(WMT), leads to the same test outcomes for all the pairwise comparisons. This observation matters for the reproducibility comparison study between the t-test and the WMT as the result is near enough the same.

4.3 NPI-B-RP for pairwise t-test

This section studies reproducibility for the Student's t-test for comparison of two groups from the NPI perspective, with the NPI bootstrap implementation. First, Section 4.3.1 introduces an algorithm for calculating NPI bootstrapped reproducibility (NPI-B-RP) for the t-test for comparison of two groups of data, i.e. Algorithm 5. Secondly, Section 4.3.2 presents the results of simulation studies to investigate reproducibility of the t-test both under H_0 and H_1. Following the simulation study, Algorithm 5 is applied to a pre-existing preclinical test scenario, which was introduced in Section 4.2 and reproducibility of pairwise comparisons tests for this scenario is studied in Section 4.3.3. This test scenario investigates the optimal dose of a drug. Different doses of the treatment are given to members of different groups and pairwise comparisons are carried out on a recorded variable between adjacent doses. Sections 4.3.2 and 4.3.3 explore the relationship between two test statistics, namely Cohen's d and the p-value, and NPI reproducibility. This section explores the assumption that, if the original test statistic is close to the threshold value between rejection of the null hypothesis and non-rejection, then the test can be expected to be less reproducible than when the test statistic is further away from the threshold. Reproducibility of the t-test and the WMT is also briefly compared.

4.3.1 Algorithm

Algorithm 5 uses NPI bootstrap, introduced in Section 2.3.3, to derive reproducibility probability for the t-test, indicated by NPI-B-RP. As these values result from the use of the NPI bootstrap method, they are effectively estimates. The inputs into Algorithm 5 are the two original samples, x and y, their corresponding sample sizes n_x and n_y, the number of runs h and the number of bootstrapped samples per run N. Algorithm 5 is applied with $N = 1000$ and $h = 100$.

BinHimd [31] briefly investigated NPI-B-RP for the WMT. The goal of it in BinHimd's

Algorithm 5 Calculating NPI-B-RP for the t-test for comparison of two groups

1: Apply the t-test on the two original samples, x and y, and record the test outcome: $t^* = 1$ if H_0 is rejected and $t^* = 0$ if H_0 is not rejected.
2: Draw an NPI-B sample of size n_x from sample x and an NPI-B sample of size n_y from sample y. Apply the t-test to these two bootstrapped samples.
3: In total perform Step 2 N times for $j = 1, \ldots, N$ and each time record the test outcome: $t_{B_j} = 1$ if H_0 is rejected and $t_{B_j} = 0$ if H_0 is not rejected.
4: Calculate rp, where $rp = (\sum_{j=1}^{N} \mathbb{1}_{(t_{B_j}=t^*)})/N$
5: Perform Steps 2-4 h times, denote the resulting values rp by $rp_1, rp_2, \ldots, rp_h$.

work [31, 53] was to show that NPI-B-RP provides results in line with the theoretical values for lower and upper values of NPI-RP which can be computed for relatively small sample sizes. It is not possible to calculate precise values of lower and upper reproducibility probabilities for the t-test as the t-test statistic does not monotonically increase as a function of its input data. The algorithm for calculating NPI-B-RP for the t-test has been adopted from the NPI-B-RP for the Wilcoxon Mann-Whitney test (WMT), which was presented in BinHimd's thesis [31]. The differences between this algorithm and BinHimd's are that BinHimd presented this for the WMT only and she used h=60 whereas this thesis uses $h = 100$. Another difference between Algorithm 5 and BinHimd's algorithm [31] is that Algorithm 5 defines how to report NPI-B-RP, i.e. via different statistics (min, mean, max). Moreover, different range choices are explored and the conclusions for those are presented in Section 4.3.3. This has not been addressed by BinHimd [31]. Furthermore, BinHimd did not consider further aspects of reproducibility probability, such as its relationship to other test statistics, and the application of the algorithm in real-life settings.

The reasoning behind the choices of the presentation of the algorithm outputs and the selection of values of N and h deserve an elaboration. Various summary statistics were explored via simple simulations: min, mean, median, max, the 5th and the 95th percentile, i.e. the bootstrapped 90% confidence interval, of $rp_1, rp_2, \ldots, rp_h$. There was no added value of using the 90% confidence interval together with the max and min value as there was not a big difference between the min value and the 5th percentile, or between the 95th percentile and the max value. Similarly, there was no added value of reporting

median as the mean and median for $rp_1, rp_2, \ldots, rp_h$ are very similar. The mean value of the outputs is considered to be the best indication of NPI reproducibility, this thesis also refers to this mean as the NPI-B-RP value. Min and max value of $rp_1, rp_2, \ldots, rp_h$ is reported alongside the NPI-B-RP value.

This chapter applied Algorithm 5 with $N = 1000$ and $h = 100$, but different values for h and N have been explored as well. The key goal was to achieve a balance between the computation time and accuracy. The h is set to 100 because this work sought to use a relatively large number. Increasing h from 100 to 200 or to 500 slightly widens the range between min and max of $rp_1, rp_2, \ldots, rp_h$, however, the change is very minor and the mean value differs only in the third decimal place. Using larger h leads to larger computational time by about the same amount, without noticeably increasing the accuracy. The option of increasing the value of N was considered. When N was increased from 1,000 to 10,000, the means of $rp_1, rp_2, \ldots, rp_h$ were similar; they differed only in the third decimal, meaning that the algorithm performs well at $N = 1000$. Increasing the N decreases the difference between the minimal and maximal value of the rp_i values. This is because rp_i calculation is based on N, the number of simulations, i.e. the number of bootstrap samples generated per group per run. However, the difference is very small and increasing N also proportionally increases the computation time to calculate NPI-B-RP.

Given that bootstrap samples from one group are exchangeable, and the same applies for the other group, this should lead to rp_i being Binomially distributed. Further exploration of this is outside the scope of this thesis. Arguably, repeating Step 5 of Algorithm 5 h times provides more insight. However, it would be worth looking into how would the min and max values of h values differ from using a Binomial estimate with confidence interval instead.

4.3.2 Simulation study

This section studies reproducibility probability (NPI-B-RP) for the t-test via simulations, where reproducibility is calculated using Algorithm 5. The null hypothesis is $H_0 : \mu_x = \mu_y$ and the alternative hypothesis is $H_1 : \mu_x > \mu_y$, the level of significance is $\alpha = 0.05$. Data were simulated both under H_0 and under H_1. Under H_0 original data were generated from the Normal distribution with mean 0 and standard deviation 1 for both groups. Under

H_1 data were generated from two Normal distributions with different means, $\mu_x = 1$ and $\mu_y = 0$, but both with standard deviation 1. Further simulations were performed for different values of the means and standard deviations under H_1, these all led to similar results as for the case presented here.

The inputs for the simulation study are as follows: the sample size $n = 6, 10, 20$; means μ_x, μ_y and standard deviations σ_x and σ_y are as given in the previous paragraph; and the number of runs per simulation $N = 200$. For each run, one sample of size n is generated from each of these Normal distributions, the t-test is performed on these two samples and the p-value is computed, and NPI-B-RP for the t-test is calculated using Algorithm 5. Cohen's d, introduced in Section 1.3, has also been considered for the tests.

First, the relationship between NPI-B-RP and the p-value for the t-test is examined in the simulations. Figure 4.2, simulations under H_0, and Figure 4.3, simulations under H_1, display plots of these metrics for the three different sample sizes, with separate plots for the rejection cases only, i.e. p-value less than 0.05. It is clear that, as expected, reproducibility is the lowest close to the test threshold, so if the p-value is close to $\alpha = 0.05$. In such cases, NPI-B-RP tends to be lower in the case of rejection (red cases in the figures) than for non-rejection (blue cases). Low values of NPI-B-RP are worrying from a practical perspective, in particular in the case H_0 is rejected with the p-value only just below the level of significance, because many experiments are explicitly designed with the aim to find evidence supporting H_1. NPI-B-RP tends to increase when the p-value moves away from $\alpha = 0.05$, which is also as expected. Similar patterns have been observed in applications of NPI reproducibility for several other test scenarios [52, 144]. For the simulations under H_1, increasing n leads to fewer cases with larger p-values, which simply results from the test becoming more powerful for larger n. As a consequence, reproducibility for most non-rejection cases for larger n becomes relatively lower compared to non-rejection cases for small n, when data are sampled under H_1.

Secondly, the relationship between NPI-B-RP and Cohen's d is explored. Figure 4.4 shows the plots of these two metrics for simulations under H_0 and H_1. In Figure 4.4 there is a V-shaped pattern: both for the rejection cases (right side of the V-shape, in red) and the non-rejection cases (left side of the V-shape, in blue), NPI reproducibility of the t-test tends to increase when Cohen's d moves away from the area where the V-shape has

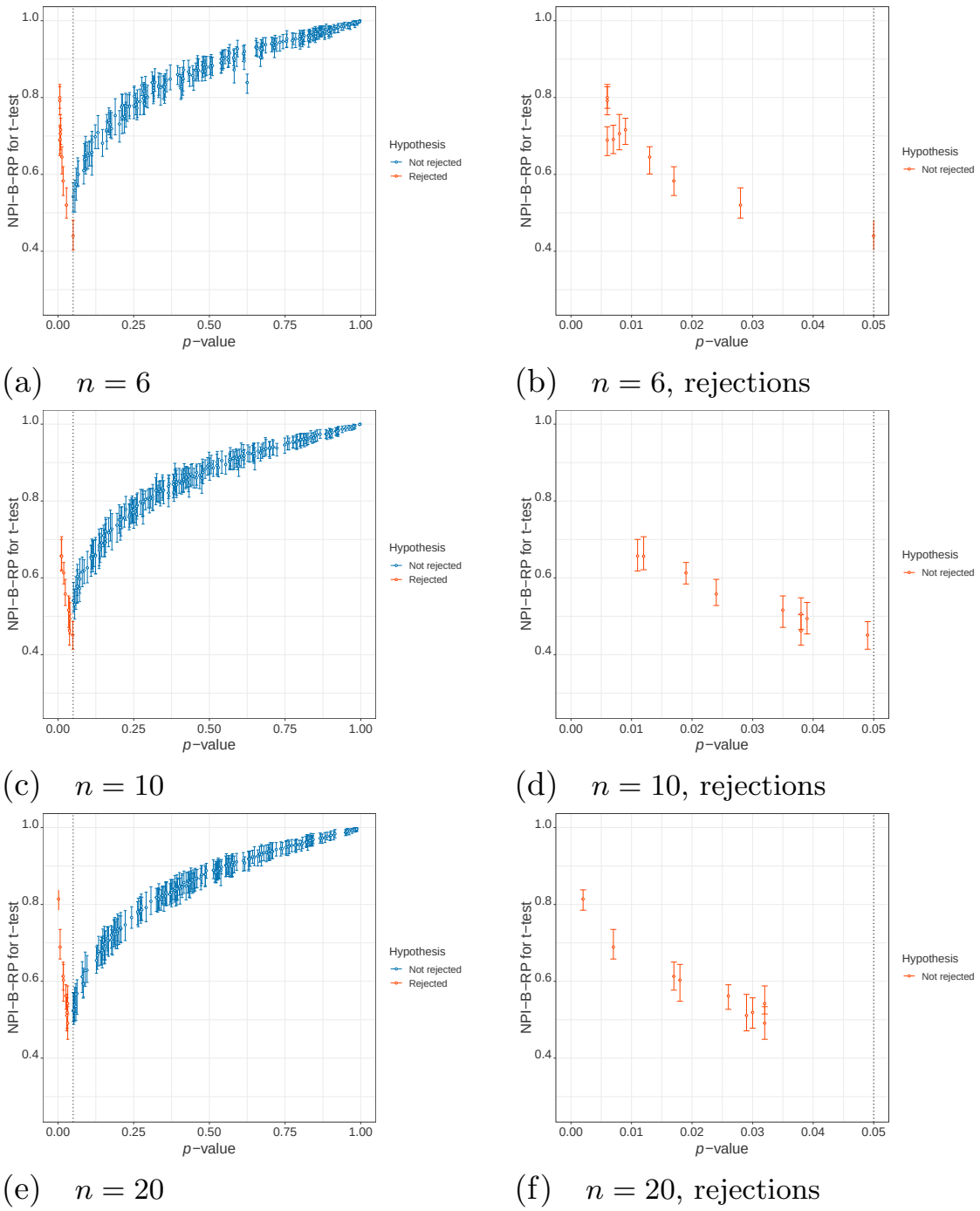
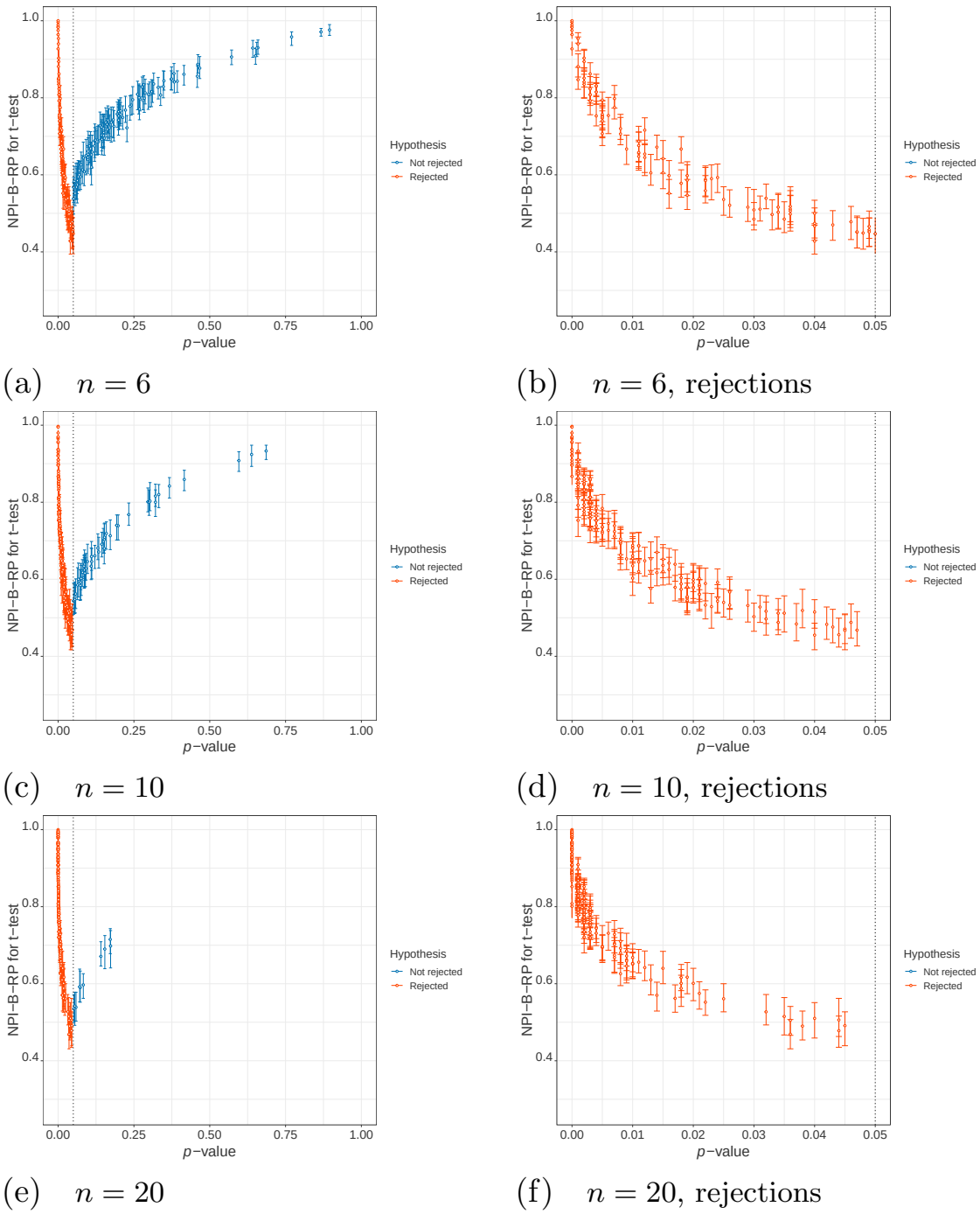


Figure 4.2: Simulations under H_0: values of NPI-B-RP (minimal, mean and maximal) for the t-test vs p-value

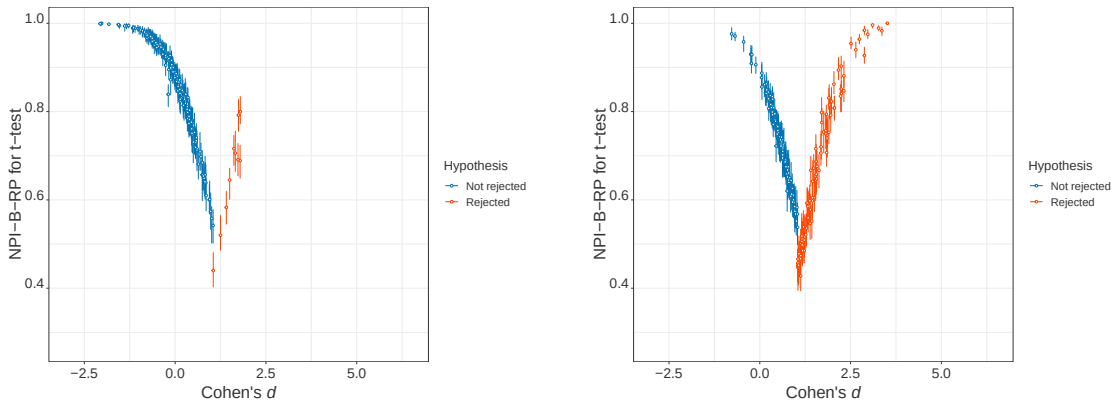


(a) $n = 6$ (b) $n = 6$, rejections

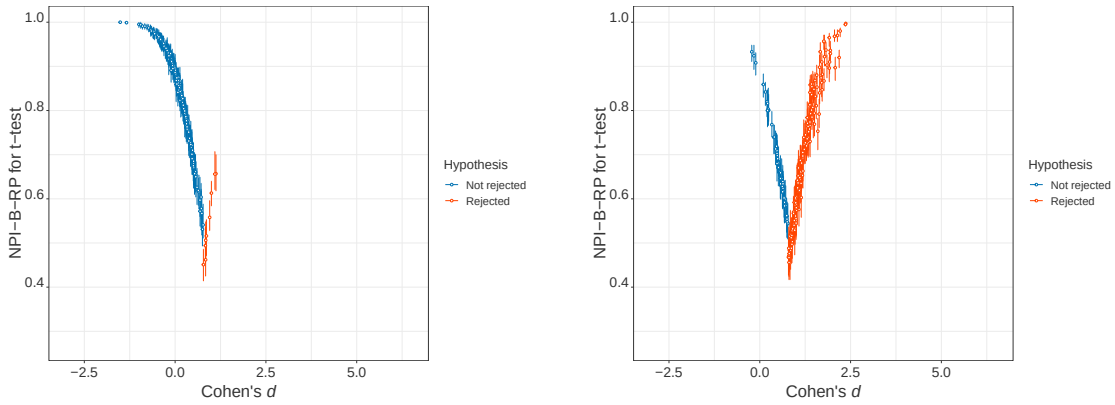
(c) $n = 10$ (d) $n = 10$, rejections

(e) $n = 20$ (f) $n = 20$, rejections

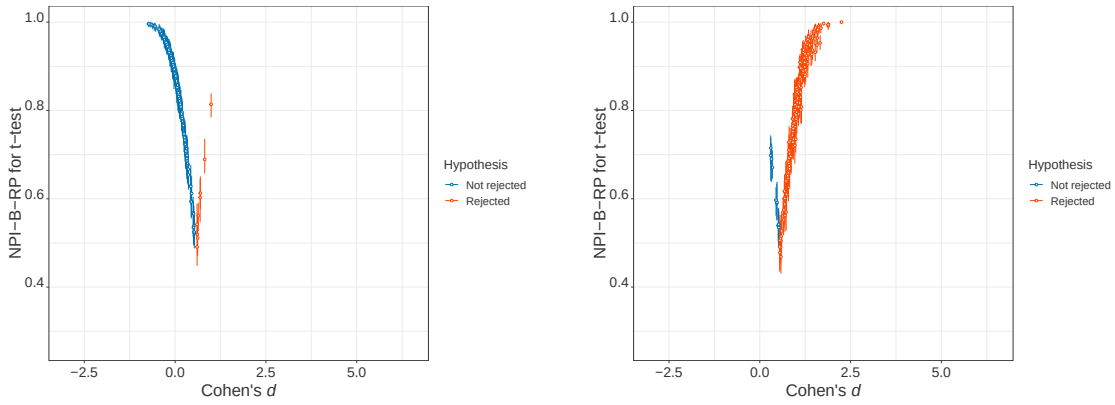
Figure 4.3: Simulations under H_1: values of NPI-B-RP (minimal, mean and maximal) for the t-test vs p-value



(a) Under H_0, $n = 6$ (b) Under H_1, $n = 6$



(c) Under H_0, $n = 10$ (d) Under H_1, $n = 10$



(e) Under H_0, $n = 20$ (f) Under H_1, $n = 20$

Figure 4.4: Simulations under H_0 and H_1: values of NPI-B-RP (minimal, mean and maximal) for the t-test vs Cohen's d

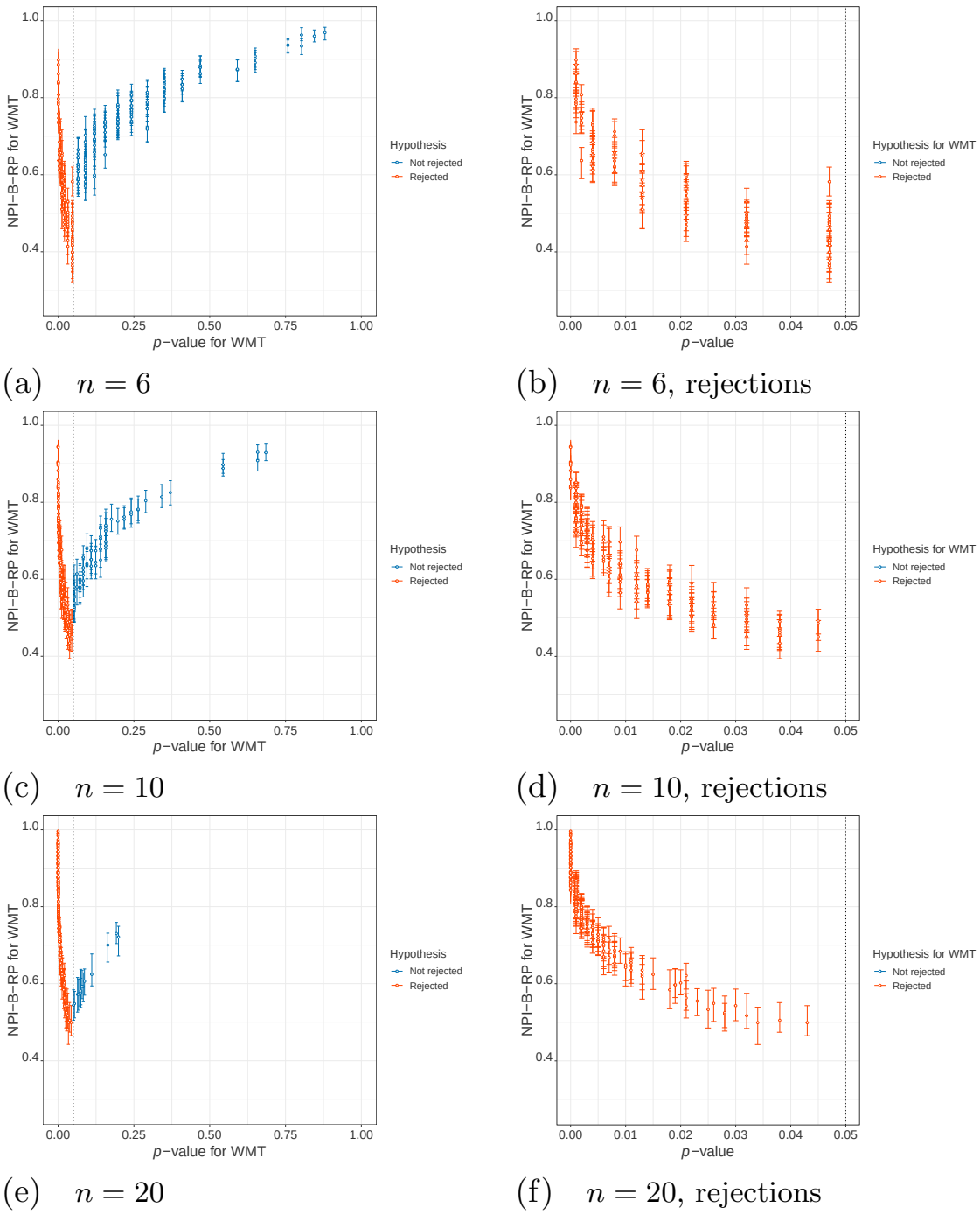


Figure 4.5: Simulations under H_1: values of NPI-B-RP (minimal, mean and maximal) for the WMT vs p-value

the lowest point. The patterns are similar across the different distribution parameters and sample sizes, where the range of the values of Cohen's d becomes a bit smaller for larger sample sizes due to the reduced variability of the sample means. The bottom of the V-shape shifts to the left, as the sample size n increases. Also, the steepness of the two legs (rejection and non-rejection upward facing lines) increases as n increases. This can be due to the fact that for large sample sizes, the range of Cohen's d tends to be smaller, possibly because bigger sample size is more representative of the population.

We ran the simulations under H_0 again, this time with $N = 1000$. The conclusions about NPI-B-RP remain the same. One thing becomes more apparent: for smaller sample size, there is bigger variety of NPI-B-RP, i.e. for the same p-value or the same Cohen's d value, NPI-B-RP differs one data set to another data set. As n gets bigger, n=20, the imaginary reproducibility probability curve gets smoother.

Finally, the variability of NPI-B-RP is studied by repeating Algorithm 5 several times for the same two data sets. Considering the mean statistics of NPI-B-RP, the repeated results differ only in the third decimal. This shows that there is low variability in the results, which means high accuracy.

It is of interest to compare NPI-B-RP for the t-test with NPI-B-RP for the Wilcoxon Mann-Whitney test [108], a frequently used nonparametric counterpart to the t-test. This is straightforward by replacing the t-test by the WMT in Algorithm 5. Figure 4.5 presents plots displaying NPI-B-RP versus the p-values for the WMT for simulations under H_1. These show a similar relationship between reproducibility probability and the p-value as for the t-test, with however fewer different p-values being possible due to the WMT, by the non-parametric nature of this test, utilising rank positions, which leads to a limited number of possible outcomes. Comparison of reproducibility of these two tests with simulated data under H_0 also led to very similar results, these are not reported here.

4.3.3 Application example

This section presents the application of NPI-B-RP for pairwise t-tests, as presented in Section 4.3.1, to a pre-existing data set from an internal preclinical study assessing the optimal dose of a drug, introduced in Section 4.2. No new experiments were carried out and the original statistical analysis framework for the experiment was adopted.

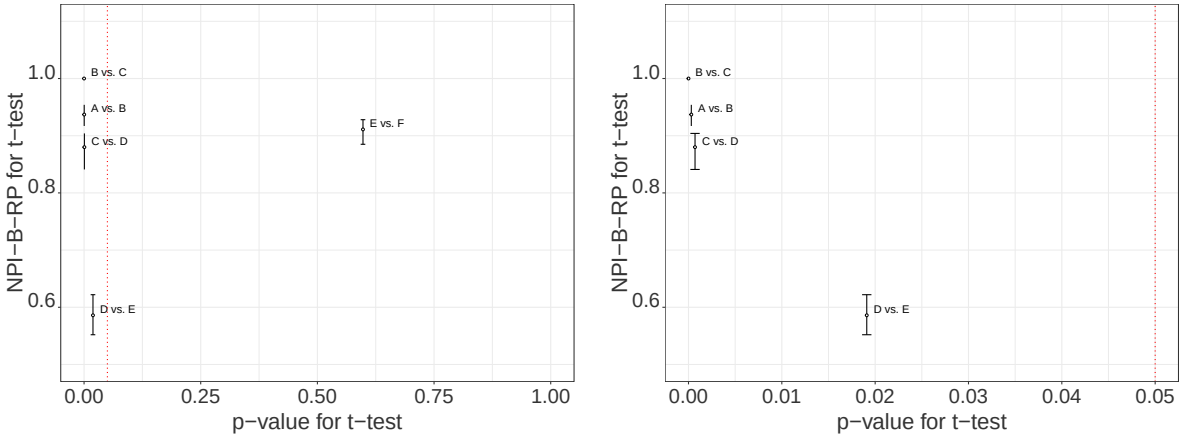
Pairwise	Statistics of the original data				NPI-B-RP					
	Reject	p-value	Effect	Cohen's	t-test			WMT		
	?		Size	d	min	mean	max	min	mean	max
A vs. B	Yes	0.0003	0.226	2.041	0.917	0.937	0.954	0.882	0.902	0.927
B vs. C	Yes	0.0000	0.366	3.213	0.999	1.000	1.000	0.999	1.000	1.000
C vs. D	Yes	0.0007	0.178	1.753	0.841	0.880	0.904	0.821	0.862	0.890
D vs. E	Yes	0.0191	0.097	1.038	0.552	0.586	0.622	0.566	0.606	0.642
E vs. F	No	0.5977	-0.013	-0.115	0.885	0.911	0.928	0.917	0.935	0.958

Table 4.2: Statistical and reproducibility analysis for pairwise comparisons

In this section, the Algorithm 5, from Section 4.3.1, is applied to the test scenario described in Section 4.2 and conclusions regarding reproducibility are drawn. The Algorithm 5 outputs and the statistics of the original test for all pairwise comparisons (A vs. B, B vs. C, C vs. D and E vs. F) are presented in Table 4.2 for finite bootstrap (Approach I, Section 2.3.3).

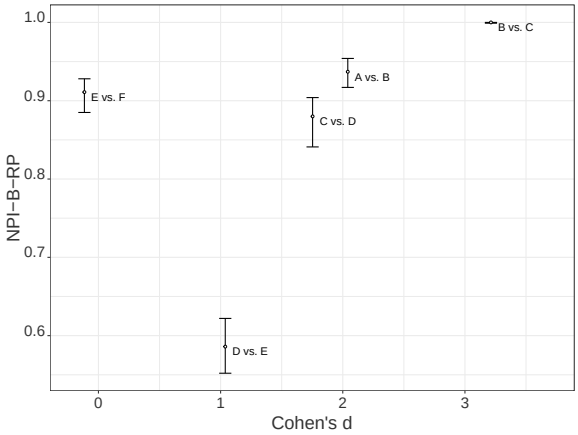
First, this work considers what conclusions about NPI-B-RP can be directly made from the preclinical test scenario. The pairwise comparison E vs. F has high NPI-B-RP value, 0.911. This means that if the test was repeated under identical circumstances and with same sample sizes, then the same test outcome would be reached with estimated probability 0.911. By comparison, the NPI-B-RP value for the pairwise comparison D vs. E is 0.586. It is up to the decision makers to consider the NPI-B-RP values alongside other statistical information and inferences, such as the effect size and power, in order to decide on the trustworthiness of the test results.

Secondly, this section explores how NPI-B-RP relates to the statistics of the t-test applied to the original data, these statistics are also displayed in Table 4.2. Note that these tests are pairwise comparisons where it is not yet taken into account that multiple tests are performed simultaneously. The Effect Size is the difference between the respective sample means; as Cohen's d is closely related to it, and the relationships between NPI-B-RP for the t-test and either the Effect Size or Cohen's d are very similar; therefore, only Cohen's d is considered in the following discussion. Figure 4.6 illustrates the relationship between NPI-B-RP for the t-test, indicating the minimum, mean and maximum values



(a) NPI-B-RP versus p-value

(b) NPI-B-RP versus p-value (rejection cases)



(c) NPI-B-RP versus Cohen's d

Figure 4.6: Comparing values of NPI-B-RP (minimal, mean and maximal) for the t-test to the statistics of the original test

of the NPI-B-RP output of Algorithm 5, for each of pairwise comparisons, and the p-values and Cohen's d. There are some clear patterns: For example, NPI-B-RP is smallest for the pairwise comparison D vs. E, where the p-value is closest to the threshold value 0.05 and Cohen's d is small. A further observation is that high NPI-B-RP values are obtained for several of the pairwise comparisons, both for some cases where the null hypothesis is rejected, in particular for the comparison B vs. C, and for the comparison E vs. F where the null hypothesis is not rejected. For B vs. C, the p-value is very small compared to $\alpha = 0.05$ and Cohen's d is very large, as Cohen's d greater than 0.8 is typically considered to be large [43]. For E vs. F, the p-value is very large compared to $\alpha = 0.05$ and Cohen's d is negative. This thesis concludes that the observations about NPI-B-RP for the preclinical test scenario are consistent with the observations made in

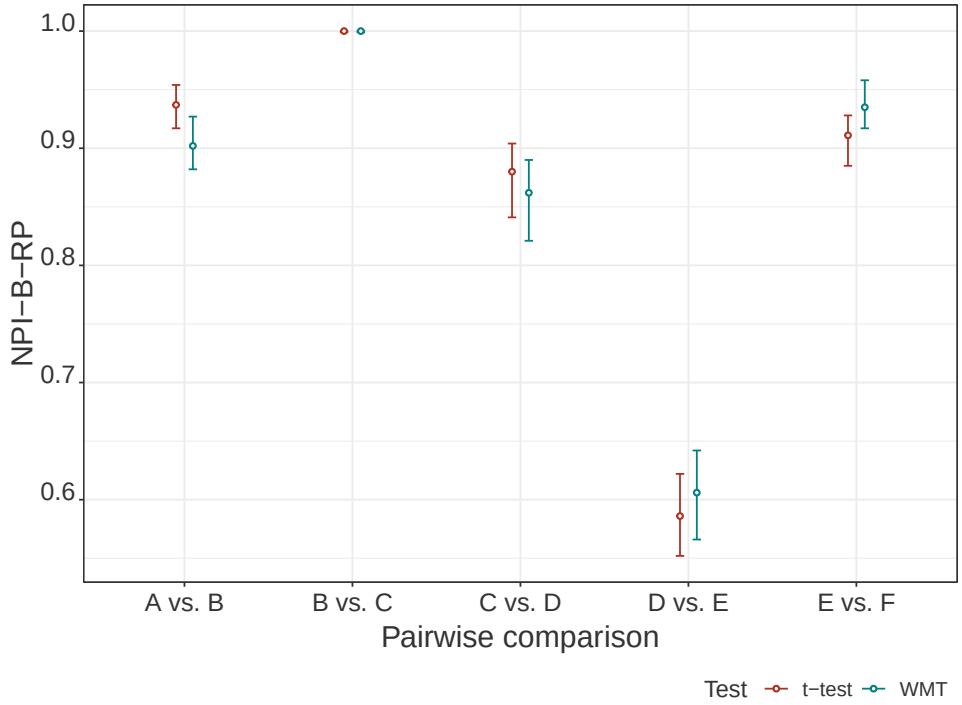
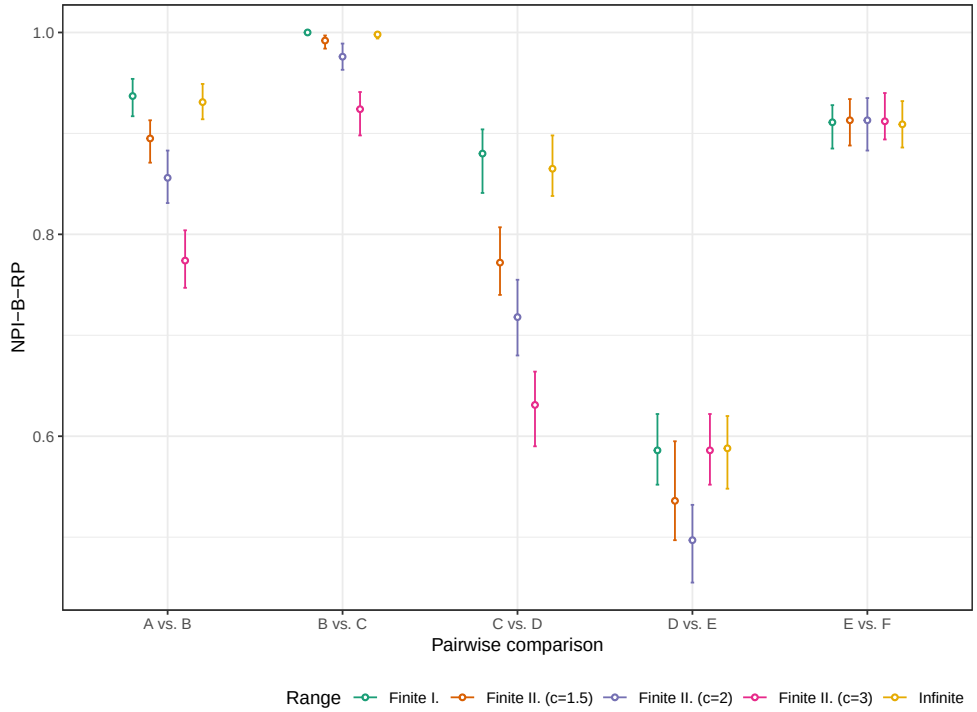


Figure 4.7: Values of NPI-B-RP (minimal, mean and maximal) for the t-test and the WMT

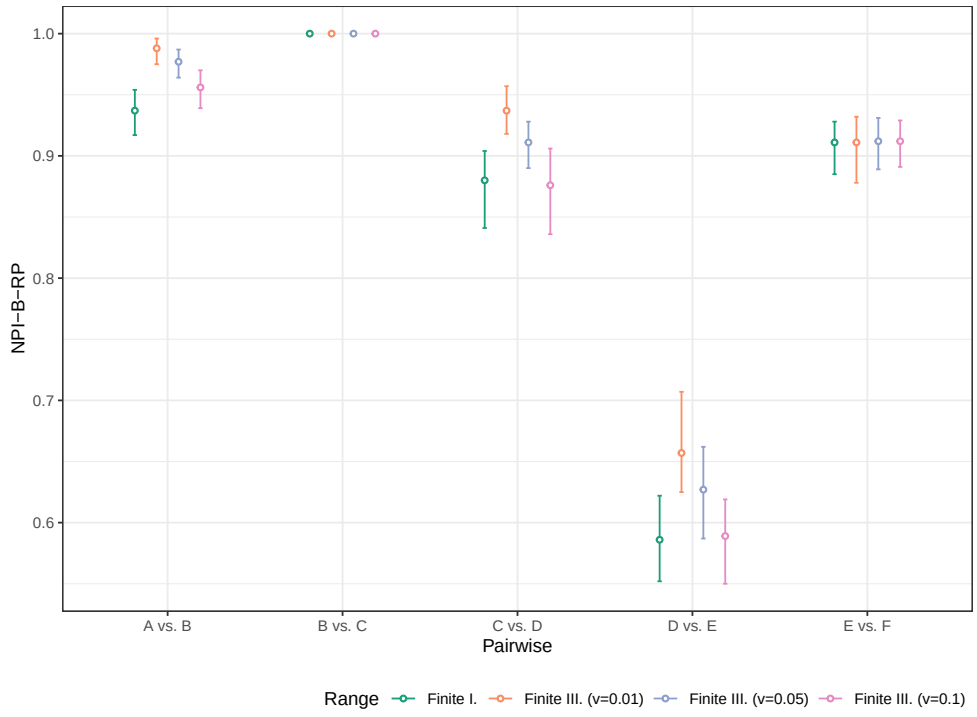
Section 4.3.2. The key observations are: NPI-B-RP is low when the p-value is close to the level of significance α. For non-rejection cases, even when the p-value is much greater than α, NPI-B-RP stays a bit below 1.

Thirdly, NPI-B-RP for the t-test and for the WMT are compared (Figure 4.7). It can be inferred that there is a pattern between NPI-B-RP for the WMT and the t-test from Table 4.2. NPI-B-RP for both tests for this case study are quite similar. This may be due to the log transformed data being an approximately Normally distributed data set. This conclusion agrees with the conclusions made in the simulation study.

Lastly, we examine how the impact of the range choice in the algorithm implementation upon reproducibility is examined. To this point, Approach I (Section 2.3.3) has been used when calculating NPI-B-RP for the t-test and the WMT. The range can be chosen as described in Section 2.3.3. NPI-B-RP for the t-test for all the approaches to the choice of range (Approach I, II, III and IV, Section 2.3.3) is illustrated in Figure 4.8. The reproducibility outcome does not change notably when using different ranges for NPI-B in Algorithm 5. There is, however, an observed pattern: If the range of the finite bootstrap



(a) Finite range (Approach I and II) and infinite range (Approach IV)



(b) Finite range (Approach I and III)

Figure 4.8: NPI-B-RP for the t-test for different range selection

is widened, reproducibility decreases. This is because there is more overlap between doses that are being compared. The largest difference between the reproducibility outcomes with different range choices can be observed in pairwise comparison between dose C and D; the cause of this difference is a topic for future exploration. As can be observed in Figure 4.8, the smallest difference between reproducibility with different range choices can be observed in the pairwise comparison between doses B and C, when there is no overlap between the data, and between doses E and F, the only pairwise comparison where the null hypothesis was not rejected. Therefore, whether the null hypothesis in the original test is rejected or not affects to what extent do NPI reproducibilities calculated through different ranges vary. The numeric results for NPI-B-RP for the studied approaches to NPI-B range choice, displayed in tables, can be found in Appendix B.1. It is outside the scope of this thesis to determine what approach of range is the best for calculating NPI-B-RP for the t-test.

4.4 Reproducibility of the final decision for the multiple pairwise comparisons

Section 4.3 introduced NPI-B-RP for the t-test for the comparison of two groups, studying the methodology via a simulation study on a preclinical test scenario introduced in Section 4.3.3. However, in this test scenario, the final choice of a particular dose is based on the multiple pairwise comparisons. This section explores NPI-B-RP of this final decision and presents a general algorithm for calculating such reproducibility. This final decision is of interest as in practice decisions are often based on more than one single statistical test; hence studying its reproducibility is important and to date has received little attention in the literature.

In a case involving multiple pairwise comparison tests, it is important to consider how the final decision is made, and which dose is finally selected pairwise. This chapter considers the scenario that the decision maker selects the smallest dose for which, in the pairwise comparisons above, the null hypothesis of no difference between the results for this dose and the next larger dose, is not rejected. In the presented test scenario, this leads to dose E being chosen, and only the actual outcomes of the five pairwise tests,

Algorithm 6 Calculating NPI reproducibility of the final decision

1: For each group G_i, $i = 1, \ldots, g$, generate an NPI-B sample.
2: Apply the multiple pairwise analysis to the bootstrapped g data sets. This includes the p-value adjustment using the Benjamini and Hochberg method.
3: Record the $(g-1)$ test outcomes. For example, test outcomes YYYYN mean do not reject H_0 only for the last pairwise comparison.
4: In total perform Steps 1-3 N times.
5: Create a frequency table of all the possible combinations of test outcomes recorded in Step 4.
6: Calculate RP_D, the proportion of combinations in Step 5 that lead to the same final decision as the original tests.

which can be presented as YYYYN, leads to this final decision. Section 4.4.1 presents the general algorithm for calculating reproducibility of the final decision, and this algorithm is applied to the test scenario from Section 4.2. In Section 4.4.2 the data from the test scenario is modified in order to illustrate and explore reproducibility of the final decision.

4.4.1 Algorithm and its application

Algorithm 6 presents a general step-by-step method for calculating NPI-B-RP of the final decision. The number of groups in the multiple pairwise comparison is denoted by g. Similarly to Algorithm 5, Algorithm 6 uses NPI bootstrap with finite intervals (Approach I., Section 1.4). So for each group, G_i, $i = 1, \ldots, g$, finite end points for the range of the possible values need to be selected. The sample sizes of the bootstrap samples are the same as of the original data. Reproducibility for the final decision, denoted by RP_D, is defined as the proportion of all the combined $g-1$ test outcomes leading to the same final decision as the original tests. In order to account for the fact that the five tests are run simultaneously, the p-values are adjusted for multiple testing using the Benjamini and Hochberg (BH) procedure [26] to control the false discovery rate. The Benjamini-Hochberg procedure takes a set of p-values and returns a set of H_0s to reject. For simplicity, we call the output of the BH procedure the adjusted p-values, even though these values are technically no longer p-values as explained in Storey [197]. The

Combination of test outcomes	Occurrence
YYYYY	18
YYYYN	400
YYYNY	39
YYYNN	319
YYNYY	4
YYNYN	93
YYNNY	8
YYNNN	29
NYYYN	35
NYYNY	4
NYYNN	30
NYNNN	8
NYNYN	13

Table 4.3: Frequency table (Step 5 of Algorithm 6)

adjusted p-values for each pairwise comparison for the original data are A vs. B: 0.0007; B vs. C: 2.7×10^{-6}; C vs. D: 0.0012; D vs. E: 0.0239; E vs. F: 0.5977. This procedure strives to decrease the proportion of false positives. In the test scenario, after the p-value adjustment, the test decision outcomes are still YYYYN.

Algorithm 6 is applied to the preclinical test scenario from Section 4.2 with $g = 6$ groups. N is set to $N = 1000$ and the final decision is based on the test results YYYYN, and so dose E is chosen because there is no significant indication that dose F is better than dose E. Algorithm 6 leads to two different types of outcome: A frequency table (Step 5) which provides all the combinations of test outcomes reached in N runs of Step 1-3, and the value of RP_D (Step 6), which is the proportion of all combinations of test outcomes that lead to the original test decision.

For this particular data set and final decision rule, the RP_D for an identical final decision, Step 6 of Algorithm 6, is 0.400, which is a relatively low value compared to the NPI-B-RP values for the pairwise comparisons as derived in Section 4.3.3. A more nuanced way of exploring the Algorithm 6 outputs is obtained by considering a reproducibility tree, which shows all possible combinations of the $g-1$ test outcomes occurring in the frequency

table. For the data set given in Table 4.1, there are 32 possible combinations of the five test outcomes. Not all combinations of test outcomes are generated by Algorithm 6 on this data set. Table 4.3 presents all the combinations of test outcomes and their frequencies. Figure 4.9 shows the reproducibility tree for the test scenario. The top node represents the 1000 runs of Steps 1-3 in Algorithm 6. This node splits into two nodes: Y$\cdots$, all possible test outcomes where in the first pairwise comparison the null hypothesis was rejected, each dot represents a following pairwise comparison with any possible test outcome; and N$\cdots$, all combinations of tests outcomes where in the first pairwise comparison the null hypothesis was not rejected. These branches again split, each into two, depending on the conclusion of the second pairwise comparison. For example, YY$\cdots$ means that the first and second pairwise comparisons lead to rejection of the respective null hypothesis. The same pattern is followed up to the last pairwise comparison.

The most frequent output is YYYYN, which is the same as the original test results and leads to dose E being chosen. The branch leading to this final decision is highlighted. The second most frequent output is YYYNN, leading to dose D. The fact that YYYNN is the second most frequent output can be explained by the relatively small NPI-B-RP value for the pairwise comparison between doses D and E.

Algorithm 6 is repeated, with $N = 1000$, ten times for this scenario. The resulting reproducibility trees were the same, only the numbers differed slightly, the RP_D values, so the proportion of runs leading to the same output YYYYN, were: 0.370, 0.376, 0.388, 0.400, 0.402, 0.403, 0.410, 0.412, 0.415, 0.424. By comparison, the NPI-B-RP values calculated on different separate runs of Algorithm 5 differ in the third decimal. Although small, the variability in these reproducibility probabilities is larger than for the individual pairwise comparisons, this is due to the use of multiple pairwise comparisons to determine reproducibility of the final decision.

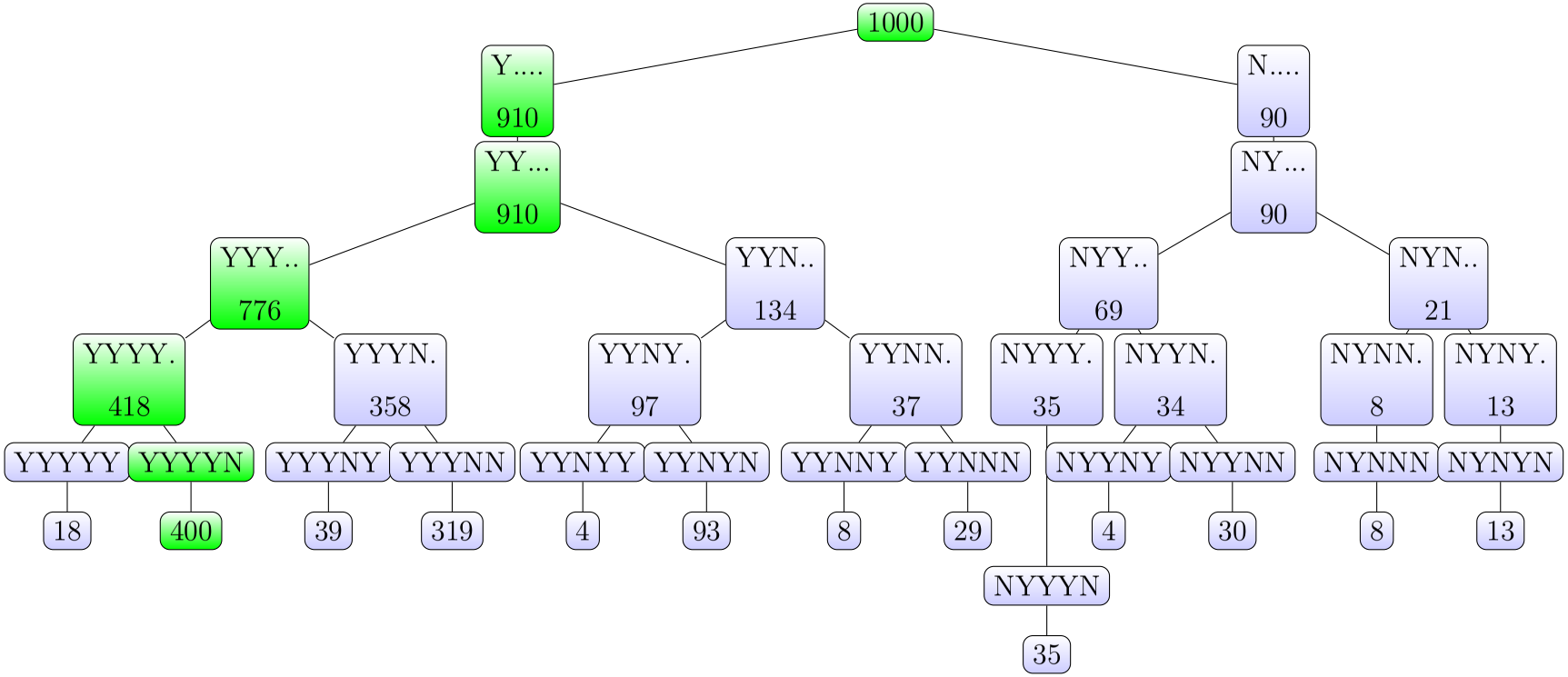


Figure 4.9: Tree diagram for reproducibility of the final decision for original test scenario (Outputs of Step 5 of Algorithm 6), finite range (Approach I)

The impact of altering the N simulations was also studied. Increasing $N = 1000$ to $N = 10,000$, increases the computing time proportionally. When N is increased to $N = 10,000$, RP_Ds calculated on 10 different separate runs of Algorithm 6 were: 0.4086, 0.3944, 0.3968, 0.4038, 0.3980, 0.3851, 0.3960 and 0.3964. Reproducibility is less varied when $N = 10000$ but the improvement is not 10 times better. Moreover, the three most common combinations of test outcomes were the same when both $N = 1000$ and $N = 10,000$ were used, this means that the outcomes are consistent and the accuracy of the results is good at $N = 1000$.

Further the effect of the choice of range upon reproducibility of the final decision is studied. Apart from the Approach I, Approach II ($c = 1$ and $c = 0.5$), and the infinite approach were explored. As the analysis showed, the wider the range, the smaller the reproducibility of the final decision. Widening the range had the same and even larger effect on the results of Algorithm 6 as it did on the results of Algorithm 5. The explanation here is the same as before: a wider range creates a larger overlap between doses. The tree diagrams for different ranges are presented in Appendix B.2.

Reproducibility for the final decision can be also studied for the WMT, by slightly adjusting Algorithm 6, i.e. using the WMT instead of the t-test. The adjusted Algorithm 6 was applied and the three most frequently occurring combinations were the same (YYYYN, YYYNN and YYNYN). Thus, reproducibility for the final decision is similar for the WMT for this particular test scenario.

4.4.2 Further illustration

If the final decision rule is followed for the test scenario data, only one combination of the pairwise test results, namely YYYYN, leads to the choice of dose E. To better illustrate the concept of reproducibility of the final decision, the data are changed for dose D by adding 1.5 to all the data points before they are log transformed, the resulting values are denoted by D' in Table 4.1 and Figure 4.1. This leads to the pairwise test outcomes YYNYN, and the final decision would be to choose dose C, since dose D does not do better than dose C. To determine reproducibility of the final decision, Algorithm 6 is again applied to the test scenario with these modified data, as is shown in Figure 4.10. Now there are 4 combinations of test outcomes that lead to the same final decision to

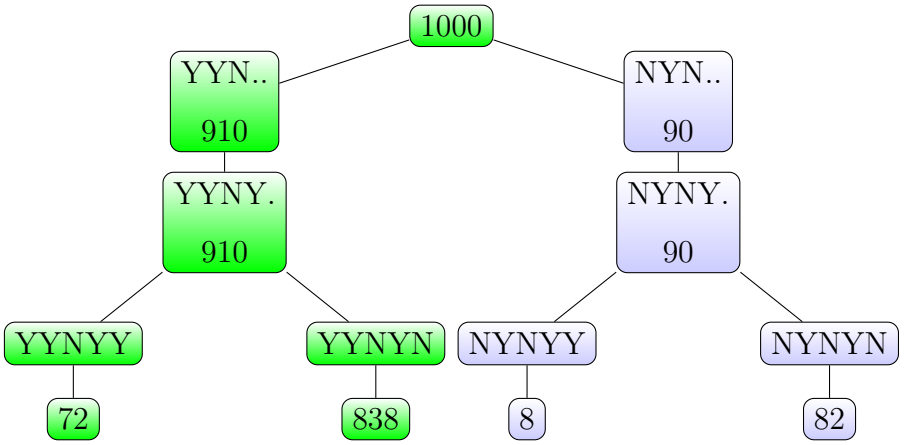


Figure 4.10: Illustration of the final decision rule: Tree diagram for reproducibility of the final decision for the modified data (Outputs of Step 5 of Algorithm 6), finite range (Approach I)

choose dose C: YYNYN (the original test outcome), YYNYY, YYNNY and YYNNN. Reproducibility of the final decision is derived as the proportion of all simulation runs in which one of these 4 combinations of test outcomes occurs. As the combinations YYNNY and YYNNN did not occur, reproducibility of the final decision for the modified data is derived by summing the proportions of runs with outcomes YYNYY and YYNYN, leading to 0.910, as highlighted in Figure 4.10. This simulation was also repeated ten times, and the results were very similar, with RP_D values 0.894, 0.910, 0.911, 0.917, 0.917, 0.917, 0.919, 0.919, 0.919, 0.922. In all these simulations, the resulting reproducibility trees were the same, with only small differences in the numbers.

4.5 Reproducibility of the WMT via the sampling of orderings

This section presents an algorithm for estimating lower and upper reproducibility probabilities for the WMT. As shown in Section 4.3, the NPI-B method provides a point estimate, whereas classical NPI uses the more general concept of imprecise probability to quantify uncertainty, hence leading to lower and upper reproducibility probabilities. Bin-Himd [31] presented an algorithm for analytically deriving lower and upper NPI-RP for

the WMT. This algorithm considers all the possible orderings of future observations based on the observed data. The concept of an ordering has been introduced in Section 1.4.

The number of all orderings to consider increases rapidly as the sample size increases. In order to calculate NPI-RP for a test comparing a group X of sample size m with a group Y of sample size n, $\binom{2m}{m}\binom{2n}{n}$ orderings need to be considered. For example, when $m = n = 10$, there are $\binom{20}{10}\binom{20}{10} = 670442572800 \times 670442572800 = 4.494932 \times 10^{23}$ possible orderings and the calculation of reproducibility probability becomes computationally expensive, and with larger sample sizes, it becomes unfeasible for a standard computer.

Coolen and Marques [55] carried out a research on determining estimates for NPI-RP through the sampling of orderings for the likelihood test. This section presents an algorithm for calculating estimates of lower and upper NPI reproducibility probability (NPI-RP) for the Wilcoxon Mann-Whitney test (WMT), using the sampling or orderings. Section 4.5.1 outlines how to analytically derive lower and upper NPI-RP for the WMT and how to estimate those, by considering a chosen number of orderings. Section 4.5.2 estimates lower and upper reproducibility probabilities for the WMT for pairwise comparisons presented in Section 4.2.

4.5.1 NPI-RP estimates for the Wilcoxon Mann-Whitney test

The upper-tailed two-sample Wilcoxon Mann-Whitney test (WMT) applied to the observed ordered data $x_{(1)} < x_{(2)} < \ldots < x_{(n_x)}$ and $y_{(1)} < y_{(2)} < \ldots < y_{(n_y)}$ leads to the rank sum test statistic Z, which was defined in Equation (1.5) in Section 1.3. H_0 is rejected if $Z \geq Z_\alpha$ and H_0 is not rejected if $Z < Z_\alpha$. As explained in Section 1.4, NPI can be used to make prediction inference for n_x future observations of X, i.e. $X_{(n_x+1)} < \ldots < X_{(2n_x)}$, based on $x_{(1)} < x_{(2)} < \ldots < x_{(n_x)}$ and there are $\binom{2n_x}{n_x}$ possible orderings of n_x future observations of X, all equally likely. Similarly, there are $\binom{2n_y}{n_y}$ possible orderings of n_y future observations of Y, $Y_{(n_y+1)} < \ldots < Y_{(2n_y)}$, all equally likely. In total, there are $\binom{2n_x}{n_x}\binom{2n_y}{n_y}$ possible orderings of n_x future observations of X and n_y future observations of Y, all are equally likely. No assumptions are made about where exactly within each interval the future observation will be located.

For each X ordering and Y ordering, let Z^f be the corresponding rank sum test statistic, which is a random variable associated with the WMT statistic for a randomly-

chosen ordering. Z^f cannot be expressed in precise numbers, therefore, a lower and an upper bound is specified [31]. Lets assume the H_0 was rejected when the WMT was applied to the original data. This test results is certainly reproduced for each combination of orderings for which Z^f must be larger or equal to the critical value Z_α. The test result could be possible reproduced for all combinations of orderings for which Z_f could be greater or equal to Z_f. If the original data lead to non-rejection of H_0, then the focus would be on the event $Z_f < Z$. In the case of the original test rejecting H_0, to calculate the NPI lower reproducibility probability, the number of orderings of future observations for which $Z^f \geq Z_\alpha$ must certainly hold are counted, while to calculate the corresponding NPI upper reproducibility probability, the number of orderings for which this event can hold [31] are counted. As noted in Section 1.3, the critical value Z_α can be read from tables. The calculation can be performed similarly for $Z^f < Z_\alpha$ in the case of the original test rejecting H_0.

Let $(S_1^X, \ldots, S_{n_x+1}^X)$ be the specific ordering of the n_x future X observations among the observed data and let $(S_1^Y, \ldots, S_{n_y+1}^Y)$ be the specific ordering of the n_y future Y observations. More explicitly, S_j^X is the number of the future observations in the interval $(x_{(j-1)}, x_{(j)})$ created by the observed data for $j = 1, \ldots, n_x + 1$, where the following conditions are satisfied: $S_j^X \geq 0$ and $\sum_{j=1}^{n_x+1} S_j^X = n_x$. According to assumption $A_{(n)}$, all the different orderings of such values S_j^X, for $j = 1, \ldots, n_x + 1$, are equally likely. Similar rules apply to S_j^Y. The left $(x_{(0)}, y_{(0)})$ and the right $(x_{(n_x+1)}, y_{(n_y+1)})$ bounds of support for X and Y, respectively, can be determined in the similar way as it was described for the NPI-B-RP approach (see Section 2.3.3).

Let $j_{(l)} = \max\{j : x_j < y_l\}$ for $l = 1, \ldots, n_y + 1$ and $j = 0, 1, \ldots, m$, so $x_{j(l)} < y_{(l)} < x_{j(l)+1}$. The rank of y_l is $l + j_{(l)}$. Then the equations for calculating lower and upper probabilities for Z^f are [31]:

$$\underline{Z}^f = \sum_{l=1}^{n_y+1} S_l^Y \left\{ \sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j_{(l-1)}-1} S_t^X + \frac{S_l^Y + 1}{2} \right\} \tag{4.1}$$

$$\overline{Z}^f = \sum_{l=1}^{n_y+1} S_l^Y \left\{ \sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j_{(l)}} S_t^X + \frac{S_l^Y + 1}{2} \right\} \tag{4.2}$$

Proof of Equations (4.1) and (4.1) can be found in BinHimd's thesis [31]. To calculate

Algorithm 7 Calculating NPI-RP for the WMT estimates for the rejection case through the sampling of orderings

1: Generate an ordering for X and for Y.
2: Calculate the corresponding rank sum test statistic, $\underline{Z}^f$ and $\overline{Z}^f$, for those orderings through Equations (4.1) and (4.2).
3: Repeat Steps 1 and 2 n^* times in total.
4: Divide the total sum of orderings where $Z^f \geq Z_\alpha$ by n^* to calculate estimates for lower and upper probabilities.

lower and upper probabilities, the total sum of orderings where $Z^f \geq Z_\alpha$ is divided by $\binom{2n_x}{n_x}\binom{2n_y}{n_y}$. Similarly, this can be done for the lower-tailed test. This has not been presented in BinHimd's thesis [31], but the calculation is straightforward.

The above method of calculating reproducibility probability considers all the orderings. To reduce the computer time, the orderings are sampled and $\underline{Z}^f$ and $\overline{Z}^f$ are calculated for those and their sum is divided by the number of orderings sampled. The procedure for the upper-tailed one-sided WMT for the rejection case is outlined in Algorithm 7. Note, in Step 4, for the non-rejection case, the total sum of orderings where $Z^f \leq Z_\alpha$ would be divided. For the lower-tailed WMT for the rejection case, the total sum of orderings where $Z^f \leq n_y(n_x + n_y + 1) - Z_\alpha$ would be divided by the number of orderings.

Confidence intervals (CI) come from the binomial properties; they are calculated using the standard result based on the Normal approximation, as shown in Equation (4.3) where $\hat{p}$ is the estimated value of the $\underline{RP}$ and $\overline{RP}$ and $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n^*} \tag{4.3}$$

4.5.2 Application example

Estimates of NPI-R-RP for the WMT can be calculated for data from the study described in Section 4.2. Algorithm 7 (with finite intervals, Approach I., Section 2.3.3) and its alteration for non-rejection cases was applied to each pairwise comparison with $n^* = 1000, 2000, 3000, 5000$. The outputs of the sampling of orderings for the WMT, estimates

Pairwise	NPI-	$n^* = 1000$				$n^* = 2000$			
	B-RP	$\underline{RP}$	CI	$\overline{RP}$	CI	$\underline{RP}$	CI	$\overline{RP}$	CI
A vs B	0.902	0.710	(0.682, 0.738)	0.975	(0.965, 0.985)	0.706	(0.686, 0.725)	0.979	(0.972, 0.985)
B vs C	1.000	0.999	(0.997, 1.000)	1.000	(1.000, 1.000)	0.999	(0.997, 1.000)	1.000	(1.000, 1.000)
C vs D	0.862	0.856	(0.834, 0.878)	0.992	(0.986, 0.998)	0.851	(0.835, 0.866)	0.995	(0.991, 0.998)
D vs E	0.606	0.446	(0.415, 0.477)	0.841	(0.818, 0.864)	0.444	(0.422, 0.466)	0.849	(0.833, 0.864)
E vs F	0.935	0.669	(0.640, 0.698)	0.939	(0.924, 0.954)	0.703	(0.682, 0.723)	0.942	(0.931, 0.952)

Table 4.4: NPI-RP estimates for the WMT using Algorithm 7 (independent samples, Approach I), $n^* = 1000, 2000$

of NPI-RP, together with CIs, alongside the outputs of Algorithm 5, NPI-B-RP, are displayed in Tables 4.4 and 4.5.

First, it is investigated how increasing n^* affects the Algorithm 7 outputs. For the given test scenario, the difference between NPI-RP estimates for different n^* is in the second decimal place, which is not very notable. Therefore, it can be concluded that at $n^* = 1000$ estimates of NPI-RP are very good.

Secondly, the outputs of Algorithm 7 were compared with the means of NPI-B-RP, i.e. the outputs from Algorithm 5. From Tables 4.4 and 4.5, it can be inferred that for all n^*, i.e. 1000, 2000, 3000 and 5000, and for all pairwise comparisons, the mean NPI-B-RP value lies between lower and upper reproducibility probabilities, $\underline{RP}$ and $\overline{RP}$. This consistency was expected from the theoretical perspective. NPI-B-RP was calculated for the lower-tail one-sided WMT, whereas NPI-RP estimates were calculated through the sampling of orderings for the upper-tail one-sided WMT. This study compared NPI-RP for the upper-tail one-sided WMT and the lower-tailed one-sided WMT. As expected, NPI-RP estimates for the upper-tailed one-sided WMT were consistent with NPI-RP estimates for the lower-tailed one-sided WMT.

Thirdly, the range of the imprecise probabilities for different pairwise comparisons can be assessed. *Imprecision* of NPI-RP is equal to $\overline{RP}$ minus $\underline{RP}$. The largest *imprecision* is for the fourth pairwise comparison, where the mean of NPI-B-RPs is the lowest. On the other hand, for the second pairwise comparison, the *imprecision* is the smallest. Here the mean value of NPI-B-RP for the pairwise comparison is 1.000. Therefore, the pattern is that for reproducibility close to 1, the *imprecision* is close 0, whereas for low reproducibility, the *imprecision* is larger than for high reproducibility. However, there is a need for

Pairwise	NPI-	$n^* = 3000$				$n^* = 5000$			
	B-RP	$\underline{RP}$	CI	$\overline{RP}$	CI	$\underline{RP}$	CI	$\overline{RP}$	CI
A vs B	0.902	0.709	(0.692, 0.725)	0.973	(0.967, 0.979)	0.719	(0.707, 0.732)	0.980	(0.976, 0.984)
B vs C	1.000	1.000	(0.998, 1.000)	1.000	(1.000, 1.000)	0.999	(0.999, 1.000)	1.000	(1.000, 1.000)
C vs D	0.862	0.857	(0.844, 0.869)	0.993	(0.990, 0.996)	0.854	(0.845, 0.864)	0.992	(0.989, 0.994)
D vs E	0.606	0.434	(0.417, 0.452)	0.842	(0.829, 0.855)	0.433	(0.419, 0.447)	0.838	(0.827, 0.848)
E vs F	0.935	0.685	(0.667, 0.702)	0.935	(0.926, 0.944)	0.685	(0.672, 0.698)	0.935	(0.928, 0.942)

Table 4.5: NPI-RP estimates for the WMT using Algorithm 7 (independent samples, Approach I), $n^* = 3000, 5000$

a further exploration of this hypothesis, as NPI-B-RP for the first and the last pairwise comparison is larger then for the third one, but the *imprecisions* for those are larger than for the third pairwise comparison.

4.6 Reproducibility of the t-test via the sampling of orderings

There are two reasons why lower and upper reproducibility cannot be analytically derived for the pairwise t-test. First, for larger samples it becomes computationally challenging to consider all the orderings in order to calculate NPI-RP. Secondly, computing the minimum and maximum values of the t-test statistic for m future observations with given ordering O_i is difficult, because this statistic depends both on the sample mean and variance. The aim of this section is to present possible heuristics for approximating NPI-RP lower and upper reproducibility probability for the t-test.

This thesis focuses on the one-sided two-sample t-test. For equal variance t-test, the t-value is calculated via Equation (1.1). In order to approximate lower and upper reproducibility probability, the approximation of lower and the upper bounds of the t-value need to be obtained. From Equation (1.1) it is apparent that there are two variables for which to account: the means (in the numerator) and the variances (in the denominator). Here the optimisation process for the upper-sided t-test, where we have two groups, X and Y, $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x > \mu_y$, are introduced.

To get the lower bound of the t-value, t_l, the procedure needs to minimise the numerator, i.e. minimise the difference between the means, and maximise the denominator, i.e.

maximise the standard deviations. Small $\overline{x}$, large $\overline{y}$ and large variances are needed. To get the upper bound of t, t_u, the procedure needs to do the opposite: maximise the numerator and minimise the denominator. Large $\overline{x}$, small $\overline{y}$ and small variances are needed.

Ideally, both of these two processes should be done simultaneously. This is an impossible task and, therefore, the approximation of NPI-RP for the t-test is split into two explorations: First one focuses on maximising and minimising the numerator (i.e. means) and this thesis refers to this approach as the numerator approach. The second one focuses on maximising and minimising the denominator (i.e. variances) and this thesis refers to it as the denominator approach.

The heuristics for calculating approximations of lower and upper RP, for both the numerator and denominator approach, are introduced in Section 4.6.1. Section 4.6.2 applies the algorithm to both approaches on the case scenario of the preclinical test study from Section 4.2. Section 4.6.3 summarises the initial findings and outlines a possible further work on the topic.

4.6.1 Heuristics for the methodology

The numerator approach

The objective of the first approach (the numerator approach) is to minimise the numerator to get the lower bound for the t-value, t_l, and to maximise the numerator to get the upper bound for the t-value, t_u. To calculate t_l for a given ordering of X and a given ordering of Y, all x's are put to the left bound of the interval and all y's are put to the right bound of the interval; and to calculate t_u, all x's are put to the right bound of the interval and all y's are put to the left bound of the interval. The denominator is also affected by this process, as the variance of the data for groups X and Y changes, but this change cannot be controlled.

The denominator approach

The objective of the second approach (the denominator approach) is to maximise the denominator to get the lower bound of the t-value, t_l, and minimise the denominator to get the upper bound of the t-value, t_u. To achieve this, the data points are split into half

Algorithm 8 Calculating NPI-RP approximations for the t-test

1: Apply the t-test on the two original samples, x and y, and record the test outcome: $R^* = 1$ if H_0 is rejected and $R^* = 0$ if H_0 is not rejected.

2: Sample a specific ordering of the n_x future X observations among the corresponding n_x data observations, $(S_1^X, \ldots, S_{n_x+1}^X)$, and a specific ordering of the n_y future Y observations among the corresponding n_y data observations, $(S_1^Y, \ldots, S_{n_y+1}^Y)$.

3: Calculate t_l and t_u for these orderings following the rules for either the numerator or the denominator approach.

4: In total perform Steps 2 and 3 n^* times for $j = 1, \ldots, n^*$ to get n^* t_l's and t_u's and each time record the test outcome for the lower and upper bound: $R_{l_j} = 1$ / $R_{u_j} = 1$ if H_0 is rejected and $R_{l_j} = 0$ / $R_{u_j} = 0$ if H_0 is not rejected.

5: Calculate the mean of t_l's, $\overline{t_l}$, and the mean of t_u's, $\overline{t_u}$.

6: Calculate lower and upper bounds for reproducibility probability: $rp_{l_j} = (\sum_{j=1}^{N} \mathbb{1}_{(R_{l_j}=R^*)})/N$ and $rp_{u_j} = (\sum_{j=1}^{N} \mathbb{1}_{(R_{l_u}=R^*)})/N$.

(and the median divides these two groups). To obtain t_l, large variance is needed. The variance can be maximised by minimising the points in the left group and maximising the points in the right group. This makes the data points more spread. To obtain t_u, small variance is needed. The variance can be minimised by maximising the points in the left group and minimising the points in the right group. This makes the points less spread.

Algorithm

Algorithm 8 presents the methodology for calculating the reproducibility measure (approximates for NPI-RP) through the sampling of orderings for the upper-sided equal variance t-test, comparing two groups, X and Y, $H_0 : \mu_x = \mu_y$ vs $H_1 : \mu_x > \mu_y$. The algorithm can be applied for both the numerator and denominator approach. In this algorithm, n^* stands for the number of sampled orderings and $x_{(0)}$, $x_{(n_x+1)}$, $y_{(0)}$ and $y_{(n_y+1)}$ are determined using the finite Approach I (Section 2.3.3).

Algorithm 9 Calculating bootstrapped t-value through NPI-B

1: Generate an NPI bootstrap for X of sample size n_x and for Y of sample size n_y (using finite Approach I, Section 2.3.3).
2: Calculate t_B for the bootstrapped samples for X and Y.
3: Do this n^* times in total to get n^* t_B values.
4: Calculate the average of the n^* t_B values, $\overline{t_B}$.

4.6.2 Application example

Estimates of the test statistic

Before considering estimates for lower and upper reproducibility probability, estimates of the lower and upper bounds for the t-values for each set of orderings, i.e. t_l and t_u, are considered for pairwise comparisons from the test scenario presented in Section 4.2. To study t_l and t_u, the average bootstrapped t-value, $\overline{t_B}$, is calculated via Algorithm 9, which is then compared to $\overline{t_l}$ and $\overline{t_u}$, calculated in Step 5 of Algorithm 8. For both procedures (Algorithm 8 and Algorithm 9), n^* is set to be $10,000$. The lower and upper t-values are calculated through the sampling of orderings for both the numerator and denominator approach and the bootstrapped t-value is calculated through the NPI-B method for the data set from Section 4.2. For both the NPI-B and sampling of orderings methods, $x_{(0)}$ and $x_{(n_x+1)}$ are defined for group X, and $y_{(0)}$ and $y_{(n_y+1)}$ are defined for group Y, by using finite intervals (Approach I, Section 2.3.3). Tables 4.6 and 4.7 display $\overline{t_B}$, $\overline{t_l}$ and $\overline{t_u}$ for different pairwise comparisons for the numerator and denominator approach, respectively.

For all pairwise comparisons, the $\overline{t_B}$ lies in between the $\overline{t_l}$ and $\overline{t_u}$. This means estimates of the t-value calculated through the sampling of orderings are consistent with the bootstrapped t-values. The range between the $\overline{t_l}$ and $\overline{t_u}$ is quite wide. However, the range between the $\overline{t_l}$ and $\overline{t_u}$ is smaller for the denominator approach than for the numerator approach. It can be inferred from the visualisation of t_l's and t_u's that the data are not skewed, therefore, the evaluation can be made based on the means of t_l's and t_u's. Similar conclusions extend to the denominator approach.

Pairwise	Statistics of the real data			NPI reproducibility analysis		
	Reject?	original t	threshold t	Sampling of orderings		NPI-B
				$\overline{t_l}$	$\overline{t_u}$	$\overline{t_B}$
A vs. B	Yes	4.298	2.120	3.149	4.943	4.065
B vs. C	Yes	7.184	2.101	6.514	7.941	7.249
C vs. D	Yes	3.781	2.110	2.674	4.650	3.641
D vs. E	Yes	2.246	2.110	1.265	3.102	2.200
E vs. F	No	-0.251	2.110	-1.215	0.700	-0.205

Table 4.6: Lower and upper t-value calculated through the sampling of orderings (for the numerator approach) and the bootstrapped t-value

Pairwise	Statistics of the real data			NPI reproducibility analysis		
	Reject?	original t	threshold t	Sampling of orderings		NPI-B
				$\overline{t_l}$	$\overline{t_u}$	$\overline{t_B}$
A vs. B	Yes	4.298	2.120	3.251	5.422	4.065
B vs. C	Yes	7.184	2.101	6.112	8.745	7.249
C vs. D	Yes	3.781	2.110	3.010	4.530	3.641
D vs. E	Yes	2.246	2.110	1.727	2.888	2.200
E vs. F	No	-0.251	2.110	-0.426	-0.131	-0.205

Table 4.7: Lower and upper t-value calculated through the sampling of orderings (for the denominator approach) and the bootstrapped t-value

Approximation of NPI-RP

So far, only the t_u's and t_l's have been considered. Step 6 of Algorithm 8 calculates estimates for lower and upper reproducibility probability. The algorithm is applied to the data set visualised in Figure 4.1 for both the numerator and denominator approach. The NPI-RP approximations are presented in Table 4.8 together with the outcomes for NPI-RP for the WMT through the sampling of orderings and NPI-B-RP for the t-test. The *imprecision*, defined in Section 4.5.2, for the numerator approach is wider than for the denominator approach. The mean of NPI-B-RP lies between the $\underline{RP}$ and $\overline{RP}$ for all the pairwise comparisons for the numerator approach. For the denominator approach,

Pairwise	mean	t-test				WMT	
		numerator approach		denominator approach		sampling of orderings	
	NPI-B-RP	$\underline{RP}$	$\overline{RP}$	$\underline{RP}$	$\overline{RP}$	$\underline{RP}$	$\overline{RP}$
A vs. B	0.937	0.706	0.992	0.773	0.989	0.710	0.975
B vs. C	1.000	1.000	1.000	1.000	1.000	1.000	1.000
C vs. D	0.880	0.595	0.964	0.679	0.945	0.856	0.992
D vs. E	0.586	0.270	0.735	0.386	0.654	0.446	0.841
E vs. F	0.911	0.840	0.986	0.934	0.960	0.669	0.939

Table 4.8: Approximation of NPI-RP for the t-test calculated via the sampling of orderings for both the numerator and denominator approach t-test compared to the mean NPI-B-RP for the t-test and estimates of NPI-RP for the WMT

the mean of NPI-B-RP lies between the $\underline{RP}$ and $\overline{RP}$ for all comparisons, except for the E vs. F pairwise comparison (the only case when H_0 is not rejected). The *imprecision* is larger for NPI-RP estimates calculated through the sampling of orderings for the t-test, especially for the numerator approach, than for the WMT.

4.6.3 Summary

To conclude, this section presented a heuristic and some initial findings on estimates of reproducibility probability for the t-test, calculated through the sampling of orderings for the t-test. Two approaches were presented: the numerator and the denominator approach. This work has shown, on limited data, that the methodology is consistent with NPI-B-RP. It might be of future research interest to try a cyclical methodology: in stage 1 use the given variance of the data and only change means, and in stage 2 also adjust the variance. However, it is not clear how to carry this out. Also, it would be of interest to calculate reproducibility for the unequal variance (Welsh) t-test. This section only dealt will calculating estimates for RP for the one-sided t-test. For the two-sided t-test, the threshold value would change from α to $\alpha/2$. Here a possible route to implementation of the sampling of orderings for the two-sided t-test is suggested: One would minimise the t-value (make t-value close to 0, i.e. as small as possible), by pushing the means of the two groups as close together as possible. To maximise the t-value, one would apply twice

the one-sided t-test and consider the absolute value and then one would work with the maximal value of the two absolute values. However, further discussion of the practicalities related to calculating NPI-RP estimates for the two-sided t-test are outside the scope of this thesis.

Future research could also explore the following questions: How many orderings should be sampled? What approach (the numerator or the denominator one) is better for the one-sided t-test? Is there any relationship between NPI-RP estimates and the p-value and Cohen's d? Does the NPI-RP estimate change as sample sizes, n_x and n_y, and the number of orderings, n^*, increase? What is the relationship between NPI-RP estimates and NPI-B-RP?

4.7 Reproducibility for the rate of growth measure data

This section explores NPI reproducibility for the rate of growth measure data, a metric commonly used in preclinical research. The focus is on reproducibility for data sets that cannot be assumed to come from a Normal distribution. For such data sets, both the WMT and the Growth rate inhibition significance analysis is applied. Growth rate (GR) inhibition is a metric which assesses effect size. It is calculated as follows:

$$\text{GR inhibition } = (1 - \mu_T/\mu_C) * 100 \tag{4.1}$$

where C stands for the control group and T stands for the treatment group. The means of C and T, μ_C and μ_T, respectively, are calculated from the rate of growth summary measure, which will be introduced in Section 4.7.1. In the GR inhibition significance analysis, the interest is in GR inhibition greater than 30%. To be GR inhibition significant, both the GR inhibition needs to be greater than 30% and the p-value needs to be less than 0.05.

The reason behind the threshold of 30% is linked to the fact that preclinical studies aim to reflect the clinical study. In oncology the interest is in how a drug will affect tumour growth. Response Evaluation Criteria in Solid Tumours (RECIST) is used for the quantitative assessment of tumour burden [24]. Tumour burden refers to the size of

a tumour. In RECIST, the diameter of the lesions can be measured before the treatment and then repeatably during the treatment period. Measuring volumes is an alternative to diameter length [24]. In a clinical study, 30% reduction in the tumour burden is classified as a partial response (PR), whereas no tumour means full response [24]. Therefore, in the GR inhibition, 0% would indicate no effect, i.e. $\mu_T = \mu_C$, 100% would indicate stasis, i.e. μ_T close to zero, and value greater than 100% would indicate tumour regression, i.e. μ_T is negative.

This section explores reproducibility of the WMT and the GR inhibition significance on three data sets. The motivation behind this section is to show that reproducibility does not have to involve only tests where solely the p-values are used to make decisions. Moreover, this section investigates how statistical reproducibility behaves for tests carried out on data that cannot be assumed to follow the Normal distribution.

First, Section 4.7.1 explains what is the rate of growth measure. Secondly, Section 4.7.2 presents an algorithm for calculating reproducibility for the rate of growth inhibition significance analysis. Thirdly, Section 4.7.3 explores reproducibility for the rate of growth data on three different not Normally distributed data sets. Lastly, Section 4.7.4 summarises the conclusions.

4.7.1 Rate of growth measure

The rate of growth measure [100] is a simple, robust and quite general model. It is usually applied to data that are not Normally distributed. This measure is calculated from measurements at various time points, rather than just at the time end points. This avoids missing values and accounts for time series data, i.e. it takes into account measurements at different time points. This rate of growth approach contrasts with the traditional T/C (the ratio of tumour volume in control versus treated animals at a given time), which uses only one single measurement, i.e. the approach used to acquire data in the test scenario from Section 4.2. The rate of growth method is commonly used on real growth data whereas the T/C is used for relative tumour growth data.

The calculation of the rate of growth measure data is not relevant for the reproducibility analysis, what matters are the rate of growth measure data values. Nevertheless, to set this work within the context of preclinical research, the process of obtaining this values

will be briefly described in what follows. In the growth rate method, values are truncated because $log_{10}(0)$ is undefined and log_{10} of values around 0 changes dramatically with little change of the number. From time series data, the rate of growth summary measure is calculated as follows:

1. Tumour volume is moved from cm^3 to mm^3 space by multiplying it by 1000 as the method works with measurements in mm.

2. Tumour volumes less than 50 mm^3 are replaced with a minimum value of 50 mm^3.

3. Data are truncated, log transformed, cleaned and data points till certain day (e.g. 31) for each animal are considered.

4. Estimated growth rate measurement for each animal is calculated. It is calculated using the generalised linear model (R function: `glm`, with NA values excluded).

4.7.2 NPI reproducibility for the growth rate inhibition significance

The calculation of growth rate (GR) inhibition has been defined in Equation (4.1). To be GR inhibition significant, two criteria need to be met: the GR inhibition is greater than 30% and the p-value is less than 0.05. Let C and T denote the control and the treatment groups, respectively. Let N denote the number of simulations and let h denote the number of runs. Algorithm 10 presents a methodology for calculating NPI-B-RP for the GR inhibition significance. This work considers N in Algorithm 10 to be 1000 and h to be 100.

4.7.3 Application example

A study monitoring tumour volume over time can be classed as select or non-select depending on how the randomisation step is implemented. Which method is used depends on how fast the tumour grows. In a select study, the animals are randomised to the treatment group based on the tumour volume. In a non-select study, the animals are randomised to the treatment group based on their body weight. Section 4.2 presented a test

Algorithm 10 Calculating NPI-B-RP for the GR inhibition significance

1: Take the two original observations for C and T, calculate the GR inhibition and apply the WMT. Record whether both the p-value is less than 0.05 and the GR inhibition is greater than 30%.

2: From the original data for each group (C and T), draw an NPI-B sample (of the same sample size as the original data for that group) and calculate the GR inhibition for the bootstrap samples and also apply the WMT to the two bootstrap samples. Record whether both the GR inhibition is greater than 30% and the p-value is less than 0.05 for these bootstrap samples.

3: In total perform Step 2 N times and calculate the proportion of getting the same test decision as was reached in Step 1.

4: Perform Steps 2-3 h times.

5: Report the summary statistics (e.g. min, mean and max) of the h outcomes. The mean is the NPI-B-RP value.

scenario, which was an example of a select study where data are Normally distributed. This section explores reproducibility for three tumour non-select preclinical studies, in which the rate of growth measure data are employed. In these studies, data sets are not Normally distributed. The rate of growth data are displayed in Table 4.9 and visualised in Figure 4.11.

In Data set 1, some measurements in dose C and E are repeated. In particular, the value $1.228730e^{-17}$ appears twice in dose C and 4 times in dose E. In Table 4.9, this value is displayed as 0.00000 due to rounding. Those were the cases where the tumour regressed and because all values are truncated, the values of the rate of growth measure are the same. For those, `jitter`, a function in R, is used in order to carry out the WMT and the NPI reproducibility analysis. In Data sets 2 and 3, there are not any repeated values.

Statistical analysis

In these studies, the main question is whether there is a treatment impact on the growth rate, which is answered by comparing the treatment group to the control group. The effect size measure is the GR inhibition. If the GR inhibition threshold is satisfied, the treatment group is taken into further studies. In all three data sets, group A presents the

Data set 1

Dose	A	B	C	D	E
	0.06115	0.00218	-0.00187	-0.00057	0.00266
	0.06168	0.00838	-0.00180	0.04491	-0.00124
	0.06359	0.03550	-0.00072	0.05119	-0.00066
	0.06815	0.03708	-0.00049	0.05139	0.00000
	0.06836	0.04579	-0.00033	0.06336	0.00000
	0.06886	0.04601	0.00000	0.06502	0.00000
	0.07521	0.05962	0.00000	0.07234	0.00000
	0.08097	0.06028	0.01863	0.08593	0.00212
	0.08118	0.06477	0.05313	0.09065	0.01663
	0.08234	0.06517	0.05533	0.09407	0.07281
	0.08418	0.06586			
	0.08755	0.06796			
	0.08807	0.07084			
	0.08960	0.07895			
	0.09731	0.08062			

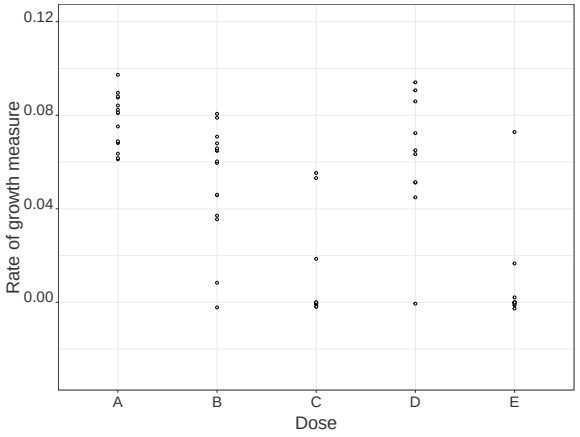
Data set 2

Dose	A	B	C	D	E	F	G
	0.04830	0.05619	0.03704	0.03826	-0.02015	-0.00768	-0.01644
	0.05694	0.05631	0.04599	0.04169	-0.01436	0.00107	-0.01349
	0.06074	0.06072	0.04631	0.05812	-0.01247	0.01555	-0.01197
	0.06583	0.06942	0.04692	0.06272	0.00000	0.02754	-0.01049
	0.06600	0.07003	0.04715	0.06826	0.01893	0.02805	-0.00037
	0.06996	0.07757	0.04985	0.06922	0.02283	0.03494	0.01648
	0.07040	0.08424	0.05161	0.08759	0.02766	0.03569	0.02167
	0.07069	0.08503	0.06293	0.08795	0.03205	0.03763	0.02591
	0.07071		0.06577	0.09466	0.03552	0.04749	0.02923
	0.07119		0.07037		0.05456	0.05163	0.02998
	0.07316		0.07234		0.06092	0.06265	0.03840
	0.07611		0.07365		0.07359	0.07819	0.04381
	0.07915						
	0.07928						
	0.08173						
	0.09530						
	0.10313						

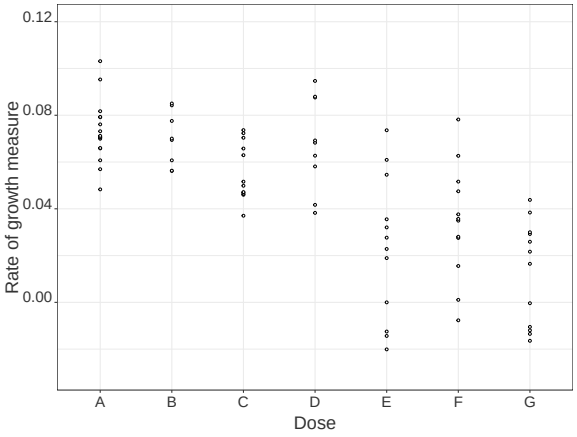
Data set 3

Dose	A	B	C	D	E	F
	0.00000	0.05291	0.03459	-0.00324	0.04540	-0.01001
	0.05248	0.06247	0.04183	0.00024	0.05036	-0.00431
	0.05842	0.06319	0.04726	0.02107	0.05517	0.03777
	0.06869	0.06540	0.06018	0.02131	0.05725	0.04258
	0.08017	0.06617	0.07081	0.03998	0.06181	0.05980
	0.08092	0.06676	0.07422	0.04858	0.07741	0.07566
	0.09237	0.06710	0.07614	0.05398	0.07842	0.08397
	0.09438	0.07158	0.07783	0.05428	0.08303	0.08570
	0.09774	0.07189	0.07785	0.06546	0.09101	0.09522
	0.09965	0.07906	0.08200	0.11677	0.10290	0.09736
	0.09988	0.08045				
	0.10311	0.08165				
	0.10574	0.08264				
	0.11570	0.09225				
	0.11796	0.09536				

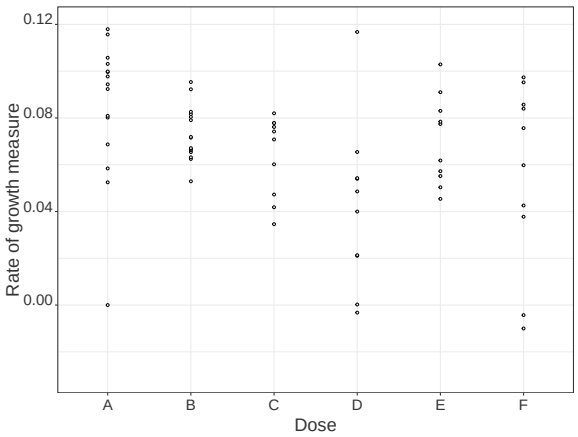
Table 4.9: Rate of growth measure data for the three data sets



(a) Data set 1



(b) Data set 2



(c) Data set 3

Figure 4.11: Rate of growth measure for the three data sets

control group. In Data sets 2 and 3, group B denotes a carrier with an inactive antibody, this is an example of a group with no effect. The rest of the groups receive treatment. For each group, there is a time series data, where the tumour volume is recorded at particular days. The rate of growth approach is applied to the data, so for each group member there is one recorded value: the rate of growth measure. Each test group is compared to the control group, i.e. group A, using the WMT. The test groups are independent compounds. So within this experiment there are multiple independent studies happening, in effect, simultaneously. Significance is assessed using a mixture of the p-value and the GR inhibition significance. To be of significant interest, the p-value needs to be below 0.05 and the GR inhibition above 30%.

Table 4.10 displays the statistical analysis for the three data sets. The aim of the tests is to determine whether the treatments work or not and if so then study them further.

Data set	Pairwise	p-value	Reject?	GR inhibition	significant GR inhibition?
1	A vs. B	0.00048	Y	32.3	Y
1	A vs. C	0.00004	Y	84.2	Y
1	A vs. D	0.17750	N	19.9	N
1	A vs. E	0.00014	Y	88.7	Y
2	A vs. B	0.54861	N	4.0	N
2	A vs. C	0.00223	Y	23.4	N
2	A vs. D	0.39581	N	7.2	N
2	A vs. E	0.00003	Y	68.1	Y
2	A vs. F	0.00003	Y	52.8	Y
2	A vs. G	0.00000	Y	82.5	Y
3	A vs. B	0.04084	Y	13.3	N
3	A vs. C	0.01628	Y	23.9	N
3	A vs. D	0.00543	Y	50.5	Y
3	A vs. E	0.05451	N	16.8	N
3	A vs. F	0.03572	Y	33.3	Y

Table 4.10: Statistical analysis of the three data sets (WMT is carried out)

For Data set 1, treatments B, D and E are GR inhibition significant, for Data set 2, treatments E, F and G are GR inhibition significant and for Data set 3, doses D and F are GR inhibition significant.

NPI reproducibility

NPI reproducibility of the statistical tests is calculated, using various methods, such as the NPI bootstrap and the NPI sampling of orderings. NPI reproducibility is calculated for the WMT and for the GR inhibition significance. The results are displayed in Tables 4.11 and 4.12, respectively. NPI-B-RP is calculated for the WMT using Algorithm 5 (through NPI-B, using finite Approach I and infinite Approach V, Section 2.3.3) and estimates of NPI-RP for the WMT are calculated using Algorithm 7 (through the sampling of orderings, using finite Approach I and n^*=1000). It can be concluded that the mean NPI-B-RP values lie between the lower and upper estimate of NPI-RP for all the pairwise comparisons (see Table 4.11), which shows consistency between estimates of NPI reproducibility. Figure 4.12 illustrates the relationship between the p-values and NPI-B-RP and estimates of NPI-RP for the WMT, respectively. The red dotted line in these figures shows the threshold value (p-value equal to 0.05). On both sides of the red line, NPI-B-RP increases the further the p-value is away from the threshold value. However, NPI-B-RP goes up more

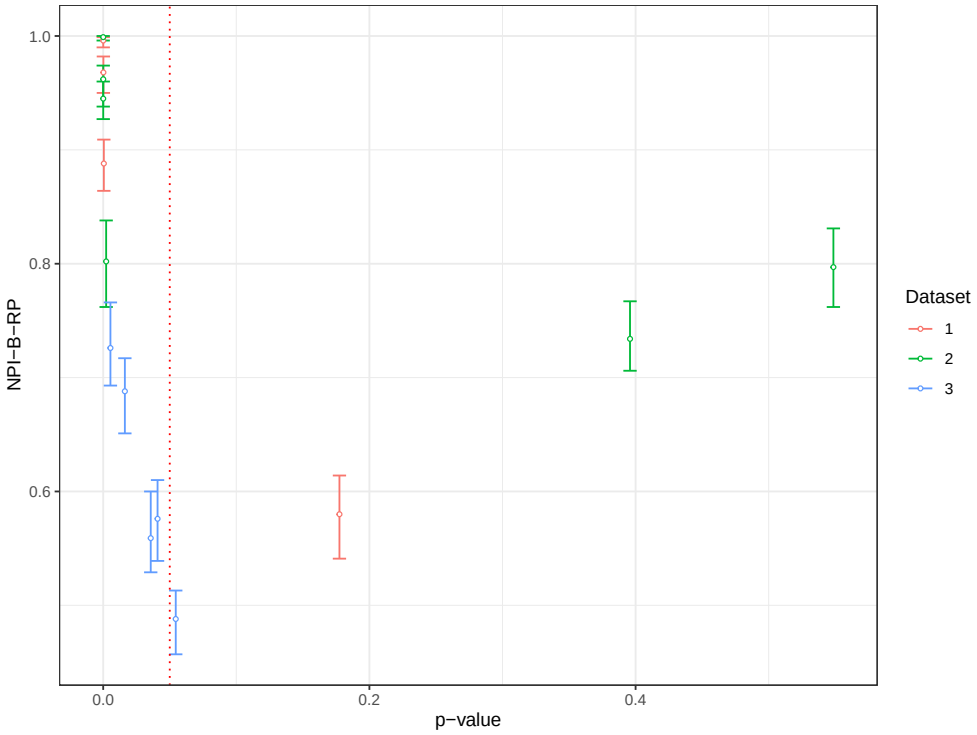
Data-set	Pairwise	p-value	Reject H_0?	Algorithm 5 (Approach I)			Algorithm 5 (Approach V)			Algorithm 7	
				min	mean	max	min	mean	max	$\underline{RP}$	$\overline{RP}$
1	A vs. B	0.00048	Y	0.863	0.885	0.913	0.864	0.888	0.909	0.807	0.968
1	A vs. C	0.00004	Y	0.987	0.993	0.999	0.990	0.996	0.999	0.952	1.000
1	A vs. D	0.17750	N	0.537	0.580	0.623	0.541	0.580	0.614	0.416	0.697
1	A vs. E	0.00014	Y	0.921	0.943	0.959	0.950	0.968	0.982	0.915	0.996
2	A vs. B	0.54861	N	0.769	0.797	0.826	0.762	0.797	0.831	0.613	0.878
2	A vs. C	0.00223	Y	0.710	0.732	0.763	0.762	0.802	0.838	0.680	0.934
2	A vs. D	0.39581	N	0.767	0.804	0.842	0.706	0.734	0.767	0.542	0.839
2	A vs. E	0.00003	Y	0.954	0.968	0.981	0.938	0.962	0.974	0.916	0.998
2	A vs. F	0.00003	Y	0.932	0.950	0.966	0.927	0.945	0.960	0.887	0.986
2	A vs. G	0.00000	Y	1.000	1.000	1.000	0.996	0.999	1.000	1.000	1.000
3	A vs. B	0.04084	Y	0.537	0.579	0.622	0.539	0.576	0.610	0.436	0.712
3	A vs. C	0.01628	Y	0.664	0.700	0.736	0.651	0.688	0.717	0.550	0.814
3	A vs. D	0.00543	Y	0.686	0.725	0.757	0.693	0.726	0.766	0.535	0.875
3	A vs. E	0.05451	N	0.444	0.487	0.521	0.457	0.488	0.513	0.275	0.659
3	A vs. F	0.03572	Y	0.509	0.546	0.586	0.529	0.559	0.600	0.394	0.781

Table 4.11: Reproducibility of the WMT for the three data sets

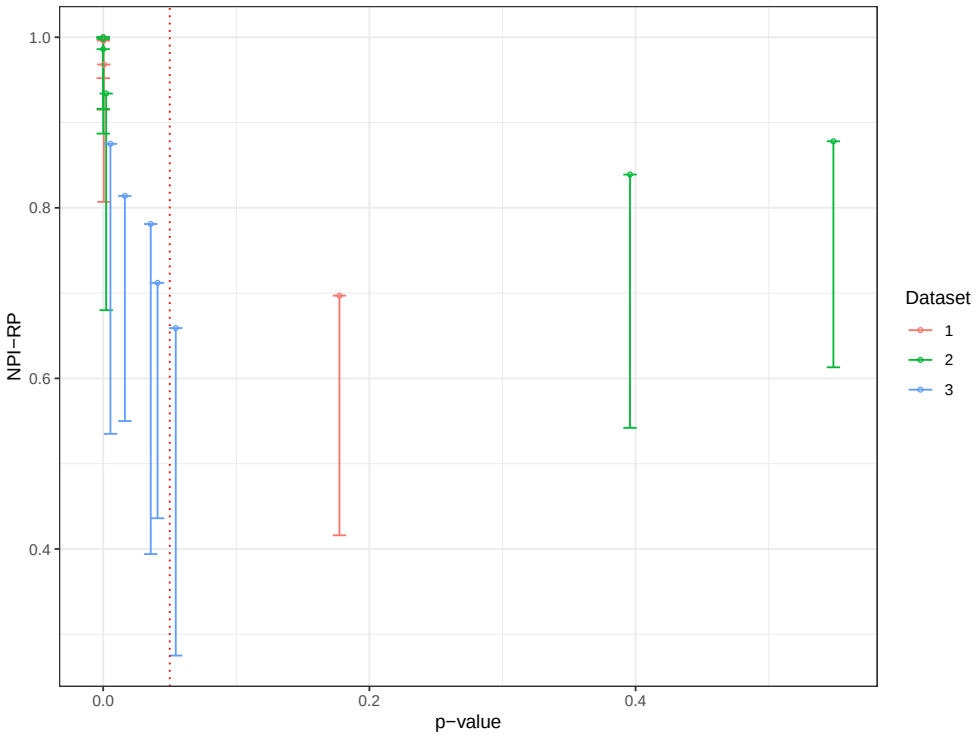
Data set	Pairwise comparison	Original GR	Significant GR?	Reproducibility for the GR significance		
				min	mean	max
1	A vs. B	32.3	Y	0.413	0.444	0.480
1	A vs. C	84.2	Y	0.975	0.983	0.992
1	A vs. D	19.9	N	1.000	1.000	1.000
1	A vs. E	88.7	Y	0.965	0.976	0.989
2	A vs. B	4.0	N	1.000	1.000	1.000
2	A vs. C	23.4	N	0.781	0.813	0.845
2	A vs. D	7.2	N	1.000	1.000	1.000
2	A vs. E	68.1	Y	0.937	0.951	0.964
2	A vs. F	52.8	Y	0.852	0.876	0.902
2	A vs. G	82.5	Y	0.981	0.992	0.998
3	A vs. B	13.3	N	0.885	0.902	0.925
3	A vs. C	23.9	N	0.624	0.661	0.708
3	A vs. D	50.5	Y	0.737	0.771	0.809
3	A vs. E	16.8	N	0.746	0.784	0.806
3	A vs. F	33.3	Y	0.470	0.519	0.557

Table 4.12: Reproducibility for the GR

steeply on the left side than on the right side. On the left side the increase is more gradual. Similar patterns were drawn for NPI-RP estimates for the WMT versus the p-value for the WMT.



(a) NPI-B-RP for the WMT versus p-value



(b) NPI-RP through the sampling of orderings for the WMT versus p-value

Figure 4.12: NPI reproducibility for the WMT versus p-value for all three data sets

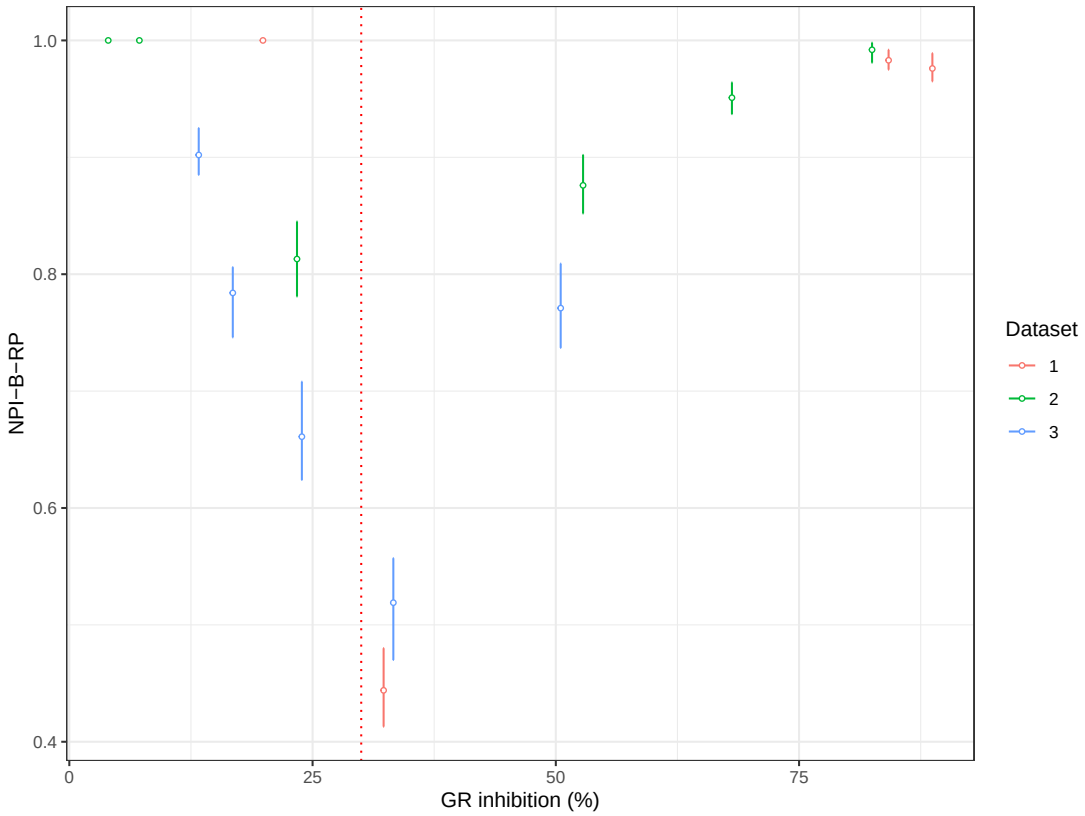


Figure 4.13: NPI-B-RP for the GR inhibition versus GR inhibition for the three data sets

Reproducibility for the GR inhibition significance is calculated using Algorithm 10. In this algorithm, NPI-B is applied on finite range (Approach I, Section 2.3.3). Similarly to the observations for NPI-B-RP for the WMT, a visible pattern between GR inhibition and NPI-B-RP can be observed in Figure 4.13. The red dotted line in this figure shows the threshold value (30%). On both sides of the red line, NPI-B-RP increases the further the GR inhibition is away from the threshold value. However, NPI-B-RP for the GR inhibition significance test goes up more steeply on the left side than it does on the right side. On the left side the increase is more gradual. Similar patterns about NPI reproducibility and the threshold test value were drawn in Section 4.3.

4.7.4 Summary

In Section 4.7.3, three data sets from preclinical studies were analysed. The goal of this investigation was to examine how reproducibility of statistical tests behaves for tests carried on data that does not follow Normal distribution and also to show that reproducibility

of statistical tests does not necessarily always need to involve solely p-values – it depends on the original statistical analysis. There is a similar prevailing pattern for both reproducibility of the GR inhibition and the p-value. A key observation was that for pairwise comparisons where the measure is close to the threshold value, NPI reproducibility goes steeply down, especially on the left side of the threshold. So for p-values close to the threshold or for GR inhibition close to the threshold, reproducibility is about 0.5.

4.8 Concluding remarks

NPI reproducibility provides an inference method for the probability of the event that, if a test was repeated under identical circumstances and with the same sample size, the same test outcome would be reached. This chapter contributed to the development of NPI reproducibility by exploring the estimation of reproducibility for the two-sample Student's t-test and for the WMT via two implementations of NPI: the NPI bootstrap the and NPI sampling of orderings.

First, in Section 4.3, NPI-B-RP for the pairwise t-test has been studied via simulations and on the application to such tests in a preclinical scenario. We explored reproducibility of the pairwise t-test and investigated the relationships between NPI reproducibility and two common test statistics, the p-value and the Cohen's d. As the p-value approaches the significance level α, NPI reproducibility decreases, and for p-values close to α NPI reproducibility is typically lower in the case of rejection of the null-hypothesis than for non-rejection. A similar pattern was seen when reproducibility and Cohen's d were compared, and further simulations, beyond the cases presented in this thesis and with other input parameters, led to similar results. Reproducibility of the t-test and the Wilcoxon-Mann-Whitney test were also compared in the simulations and for the preclinical test scenario, and the results were quite similar. This might be due to the fact that the considered data could, after transformation, reasonably be assumed to come from a Normally distributed population, and the data in the simulation study were generated from Normal distributions. A more detailed investigation of differences in reproducibilities of these two tests, for example for data from skewed distributions, is a topic for future research.

NPI reproducibility for the pairwise t-tests can provide useful insights for practical

applications. For example, in the preclinical test scenario, one of the pairwise comparisons had low reproducibility, so it might be advisable to explore the comparison of those two groups in more detail, possibly by additional experiments. NPI reproducibility can be used in conjunction with other test statistics, such as the p-value and the Cohen's d, to support the decision process based on the data and tests.

Secondly, in Section 4.4, reproducibility for a final decision based on multiple pairwise t-tests has been investigated. In the preclinical scenario considered in this chapter, multiple comparisons are performed and their test results lead to a final decision on an appropriate dose. It is, therefore, also important to consider reproducibility of this final decision; and one could say that this is the most important outcome of the combined hypothesis tests. An algorithm for deriving NPI reproducibility of this final decision was introduced, this has not been previously considered in the literature. For the presented preclinical test scenario, reproducibility of the final decision is smaller than the reproducibilities for all the pairwise comparisons on which the final decision is based. This is a logical consequence of using multiple pairwise comparisons to reach the final decision. Low reproducibility of the final decision should be taken into account by decision makers, investigating possible further actions to improve this situation is also left for future research.

Thirdly, in Section 4.5 an algorithm for calculating NPI-RP estimates for the WMT, using the sampling of orderings, is introduced and illustrated on the test scenario from Section 4.2. The methodology has been inspired by the sampling of orderings for the likelihood test [55]. A consistency between NPI-B-RP and NPI-RP estimates was shown.

Fourthly, in Section 4.6 a route to calculating NPI-RP estimates for the t-test via the sampling of orderings has been outlined. Two different approaches were introduced: one of them focuses on the numerator of the t-statistic, the other focuses on the denominator of the t-statistic. Both approaches work for the test scenario from Section 4.2: NPI-B-RP lies in between lower and upper NPI-RP estimates. However, further exploration of this methodology is left for future research.

Lastly, Section 4.7.1 studied NPI reproducibility for the WMT and GR inhibition significance test for three data sets containing the rate of growth measure data. This sections highlighted that reproducibility can also be calculated when the statistical analysis is not

solely based on the p-values. It also showed that test statistic close to the threshold value is linked to low NPI reproducibility.

This chapter concludes by highlighting ideas for future research. The main challenge is to apply NPI reproducibility to many real-world test scenarios and to use it as input into actual decision processes. Follow-up actions in the case of low reproducibility are also important and research into this has not yet been reported in the literature. Further study could investigate the sensitivity of the reproducibility calculations to the choice of the left and the right bound of the support of the bootstrap and of sampling of orderings. Moreover, it is of future research interest to study the effect of `jitter` on the WMT outputs.

Further work entails exploring the calculation of NPI-RP estimates via the sampling of orderings for the t-test via a simulation study and determining whether the numerator or the denominator approach leads to more accurate estimates of imprecise probabilities. Alternatively, the two approaches could be combined via a cyclical methodology: splitting the analysis into two stages, in Stage 1 changing only the numerator and in Stage 2, adjusting the denominator. The algorithm could also be extended to the two-sided t-test.

Chapter 5

Concluding remarks and further research topics

This thesis has made three contributions to the literature: It has shown that applying bootstrap methods provides useful inference for small samples, it has presented an overview of reproducibility and it has made advances on the topic of test reproducibility. These contributions were motivated by statistical practice in preclinical research.

This chapter aims to highlight this thesis' novel contributions to the literature and their importance in preclinical research in Section 5.1; and to outline further research topics related to small-sample bootstrap and statistical reproducibility in Section 5.2.

5.1 Summary of the findings

This section summarises the thesis contributions to small-sample bootstrap, to overview of reproducibility and to NPI reproducibility. The three topics are discussed separately in Sections 5.1.1, 5.1.2 and 5.1.3, in order to highlight the novelties of this research.

5.1.1 Contributions to small-sample bootstrap

The bootstrap method is not commonly used with small samples, as the most commonly known bootstrap method, the Efron bootstrap, is not suitable for small samples. This is because the Efron bootstrap is based on the asymptotic argument – empirical distribution, from which a sample is taken in Efron-B, converges to the real underlying population

distribution, if the number of data increases to infinity [27]. Many practitioners are not aware of the existence of other bootstrap methods or their performance with small samples. Although individual simulation or application examples of the use of the small-bootstrap method exist, these have not led to the implementation of bootstrap methods for small-sample analysis. The Banks-B method [18] has been hidden; it was introduced in the 1980s and only picked up by BinHimd [31]. This thesis has shown that Banks-B for small samples deserves further research attention. More follow-up work has been carried out for Hutson-B, as shown in Chapter 2. Nevertheless, both of these bootstrap methods are still not in the tool kit for most practitioners, due to notable insufficiency of work carried out on small-sample bootstrap.

This research therefore set out to explore via a simulation study whether a bootstrap method can provide useful inference with small samples, and to present initial recommendations on small-sample bootstrap for practitioners, as well as to identify possible areas of implementation of small-sample bootstrap. The motivation for the simulation study was driven by the common existence of small samples in preclinical research, and by the lack of the use of bootstrap methods with small samples. The main issue with small samples is that they are limited in their ability to justify model assumptions underlying most classic statistical techniques, which can increase the risk of making decisions based on wrong assumptions. The advantage of the bootstrap method is that it does not require the assumption of any underlying distribution: therefore it overcomes the problem of insecurity about the underlying distribution of small samples.

Chapter 2 compared four bootstrap methods: NPI bootstrap (NPI-B), Banks bootstrap (Banks-B), Hutson bootstrap (Hutson-B) and Efron bootstrap (Efron-B), when applied with small samples. The study provided new insights into the performance in the estimation of population characteristics (mean, variance, quantiles – Q1, median and Q3 – and IQR), and at making prediction inference for small sample sizes for these four bootstrap methods for data simulated from Normal, Lognormal, Exponential and Mixed-Normal distributions. The performance of smoothened bootstrap using Gaussian kernel (Kernel-B) is briefly addressed in Appendix A.5.

Chapter 2 concluded that Banks-B performs very well in the estimation of mean, variance and quantiles (Q1, median and Q3) with small samples ($n = 4, 6, 8, 10$), regardless

of the underlying distribution. This thesis recommends this bootstrap method for the estimation of these population characteristics for small samples. For data on the real-line, the recommendation is of a finite range (Approach I, Section 2.3.3) for Banks-B for the estimation of mean and quantiles and infinite range (Approach IV, Section 2.3.3) for the estimation of variance. For data defined on $[0,\infty)$, the half-infinite range (Approach V, Section 2.3.3) is recommended. Moreover, Hutson-B showed good performance in the estimation of quantiles for a variety of underlying distributions, and NPI-B performed well in the estimation of population characteristics for sample size $n = 4$ and in the estimation of variance for Lognormally distributed data. The study confirmed that Efron-B does not perform well in the estimation of mean, variance and quantiles for small samples. The findings of this thesis dictate a recommendation of caution to practitioners regarding the use of BC_a confidence intervals, instead of percentile confidence intervals, for the estimation of any population characteristics, before more research has been carried out.

Similarly, NPI-B performed very well at making prediction inference when predicting mean, variance and quantiles for small samples ($n = 4, 6, 8, 10, 20$). Thus, the study of the bootstrap method performance in prediction extends the conclusions of BinHimd [31] to smaller sample sizes. This means that NPI-B is a suitable method for prediction, and it can be used as a tool to calculate NPI reproducibility. This thesis would not recommend Efron-B, Banks-B or Hutson-B for prediction inference.

5.1.2 Overview on reproducibility

Chapter 3 presented an elaborate literature review on reproducibility. The issue identified in this research field was the lack of consistency in defining reproducibility and related terms. Thus, Chapter 3 began by classifying the reproducibility definitions and concepts from the literature into five Reproducibility Types: Type A to Type E. Reasons for low reproducibility and suggestions for improving reproducibility were discussed. Reproducibility in relation to preclinical research was covered, focusing on the shift from striving for homogeneity to embracing variability. The main focus of Chapter 3 was on the debates relating to statistical reproducibility. Chapter 3 highlighted the discussion of variability across studies. The focus was on how to control and quantify reproducibility and the ongoing discourse on whether to use p-values. In the latter discussion, the

conclusion reached in this thesis was that although there are many issues and problems associated with p-values, there is no clear and straightforward alternative to p-values that could be widely adopted by researchers.

Moreover, Chapter 3 outlined some of the metrics relating to scenarios where both the original and the replicate experiments have been carried out, determining whether the reproducibility has been successful. This thesis focused on quantifying statistical reproducibility in cases where only the original experiment has been carried out. A summary of available metrics for quantifying statistical reproducibility was given. Finally, a gap in the current debate was identified, i.e. the consideration about what can be inferred about reproducibility from data from the original study; and NPI reproducibility was discussed: a framework that was further developed in Chapter 4.

5.1.3 Contributions to statistical reproducibility

NPI reproducibility provides an inference method for the probability of the event that, if a test was repeated under identical circumstances with the same sample size, the same test outcome would be reached. NPI reproducibility is aimed at quantifying statistical reproducibility, it does not serve the function of recognising that an incorrect statistical method has been used, as confirmed in Appendix B.3. Chapter 4 contributed to the development of NPI reproducibility by employing two implementations of NPI (NPI bootstrap and NPI sampling of orderings) in order to estimate NPI reproducibility probability. An algorithm for calculating NPI-B-RP for the pairwise t-test was presented and studied via simulations and on an application in a preclinical scenario. This thesis showed that, for pairwise comparisons with p-value close to the significance level α, the NPI reproducibility is low, and, for p-values close to α, the NPI reproducibility is typically lower in the case of rejection of the null-hypothesis than in the case of non-rejection. The same pattern could be seen for the relationship between between NPI reproducibility and Cohen's d. The findings of this thesis thus suggest using the NPI reproducibility measure, alongside other test statistics, such as the p-value and the Cohen's d, to support the decision process based on data and tests.

This work extended the reproducibility study of pairwise comparisons to reproducibility for a final decision based on multiple pairwise t-tests, and considered a preclinical

scenario, where multiple comparisons are performed, with their test results leading to a final decision on an appropriate dose. Reproducibility of the final decision has not been extensively studied in the literature. This thesis pointed out that this reproducibility is notably lower than reproducibility for pairwise comparisons. Decision makers should take into account the low reproducibility of the final decision.

An algorithm for calculating NPI-RP estimates for the WMT, using the sampling of orderings, was presented. On a preclinical test scenario dataset, NPI-RP estimates for the WMT were studied alongside NPI-B-RP for the WMT and it was concluded that there is consistency between the two measures: NPI-B-RP lies between estimates of lower and upper reproducibility probability. Estimating NPI-RP for the t-test via the sampling of orderings is more challenging than for the WMT. This thesis outlined two different approaches to such calculation: either the focus can be on the numerator or on the denominator in the calculation of the t-value. Both approaches were explored using a preclinical test scenario: NPI-B-RP was in between the lower and upper NPI-RP estimates for both approaches.

Lastly, NPI reproducibility was studied for the rate of growth measure data. Reproducibility was calculated for three different studies. In all of these studies, the following reproducibility probabilities were calculated for the original test analysis: NPI-B-RP and NPI-RP estimates were calculated for the WMT and NPI-B-RP were calculated for the growth rate inhibition significance analysis. This section showed that reproducibility can also be calculated when the statistical analysis is not solely based on the p-values. It confirmed the previous observations: that test statistic close to the threshold value is linked to low reproducibility.

5.2 Further research suggestions

This research suggests many opportunities for further research. Section 5.2.1 focuses on further research related to small-sample bootstrap and Section 5.2.2 presents ideas for further work related to NPI reproducibility.

5.2.1 Further research related to small-sample bootstrap

The findings and methodology of this thesis could be further explored through a simulation study, which would provide more insight into small-sample bootstrap. One of the potential additional aspects could be the consideration of further distributions, e.g. Laplace, Weibull, Beta, Uniform, Gamma distributions, and more Mixed-Normal distributions.

The simulation study of small-sample bootstrap has pointed out a potential effect of heavy tails upon Hutson-B's performance in the estimation of Q3 for Lognormally and Exponentially distributed data. However, this has not been studied further. This thesis recommends a further study of the effect of a heavy tail on the performance of Hutson-B, as well as on other bootstrap methods (Banks-B and NPI-B). Such a study could be, for example, carried out on two Beta distributions. We would expect that NPI-B would be the least affected bootstrap method, given its large variability of bootstrap samples.

This thesis has explored the effect of the choice of the left and right bounds of support on the estimation performance of Banks-B and NPI-B in the small-sample bootstrap study, and on NPI-B-RP and NPI-RP estimates in the NPI reproducibility analysis. Both finite and infinite ranges were explored, concluding that the choice of range for NPI-B and Banks-B has an effect on the bootstrap method performance. This work would encourage further study into the effect of the range choice. Hutson-B has been defined only on a full infinite line $(-\infty, \infty)$ and on $[0, \infty)$. It would be of interest to consider developing Hutson-B for finite interval.

Moreover, Hutson [112] developed a sigmoidal quantile function estimator and a hybrid quantile function estimator, which were not considered in this research. It would be of interest for future research to study a variation of Hutson-B, using a different quantile function estimator, for small samples.

This thesis briefly studied the BC_a confidence intervals in the bootstrap comparison

study at bootstrap performance in estimation. The study excluded bootstrap-t confidence intervals. Further research into using BC_a confidence intervals with NPI-B and Efron-B for the estimation of IQR, and with Hutson-B and Banks-B for the estimation of Q1 and Q3 for Lognormally distributed data would be meaningful, as would a study of bootstrap-t confidence intervals.

The findings relating to the estimation and prediction of IQR are inconclusive and a future study could carry out further investigation. IQR is usually not commonly calculated for small samples. Thus, this study would be more for theoretical interest rather than for practical purposes.

In Appendix A.1, the issue of different types of quantile calculations leading to different sample statistics is briefly addressed. The type of quantile calculation has not notably affected the conclusions regarding the bootstrap methods' performance in the estimation of population characteristics. However, it would be meaningful to explore which type is the most suitable for calculating sample quantiles for small samples.

It would be of practical interest to look further into the use of bootstrap hypothesis testing with smaller sample sizes, possibly considering Banks-B as an alternative bootstrap method to Efron-B, especially where the relevant population characteristics are mean or variance. This thesis has carried out a small-scale study of using Banks-B instead of Efron-B in a bootstrap hypothesis study. The two-sided t-test is commonly used in preclinical research, as a biomarker or a safety outcome variable could change in either direction. Future study could explore bootstrap hypothesis testing for real-life test scenarios where the two-sided t-test is used. It might be of interest to explore further whether Banks-B would perform better than Efron-B in cases where the sample sizes of the original samples are unequal, and where the data comes from a variety of distributions. Examples of other applications for which Banks-B could be explored were presented in Section 2.2. These include: the power and sample size calculation, the estimation of standard errors and confidence intervals, the estimation of immune parameters, and the safety assessment in preclinical pharmacokinetics and toxicokinetics.

Lastly, although the bootstrap method does not assume a particular underlying distribution, it does assume that the data is representative of its population. With small sample sizes, there is no guarantee that this is the case. Linked to small-sample bootstrap, the

following questions arise: can small samples satisfy this criterion and - if so - how small can the sample size be to be sufficiently representative of the population? Answering these questions is beyond the scope of this thesis. This problem does not only arise for the bootstrap method, but also for any other statistical analysis of small samples. The logical recommendation would be to increase sample sizes, but in preclinical research it is often not possible, due to financial and ethical restrictions. Sometimes a smaller sample size can be a result of missing data, which is common in preclinical research. Thus a practitioner can only influence the choice of the statistical analysis. The availability of more tools can help the decision-maker to make a more informed choice. Both bootstrap and statistical reproducibility are additional measures that can aid the decision-maker.

This thesis concludes that small-sample bootstrap is a topic that has potential and it deserves more research attention. For further application of bootstrap methods on small samples, it would be beneficial to have tools to guide decision-makers in how to choose which bootstrap method is suitable for their purposes.

5.2.2 Further research related to NPI reproducibility

The main challenge is to apply NPI reproducibility to many real-world test scenarios, and to use it as an input into actual decision processes. This thesis focused on the WMT and the t-test. At the moment, NPI reproducibility can be calculated for a limited number of tests; apart from the WMT and the t-test, this includes the quantile test and the precedence test [5], likelihood ratio tests [144], the one-sample Sign Test, and the two-sample Kolmogorov-Smirnov Test [31].

In preclinical research, time progression is often recorded. In such experiments, an increase of a particular measure, such as a tumour or weight, is being recorded over time. Linear regression is often used to analyse such data. In order to be able to apply NPI reproducibility to linear regression, more research work needs to be carried out on NPI application in more dimensions. Moreover, in order to enable the calculation of NPI reproducibility for a wider spectrum of tests, more research is required on the reproducibility for data, which are integers, categorical or right-censored.

Hill's assumption $A_{(n)}$, on which NPI is based, does not allow for ties, i.e. repeated values. This problem was solved in this thesis by using the `jitter` function in R which

adds a small amount of noise to a numeric vector. Further research could explore other ways of solving this problem. For example, in the case of repeated values, a value could be sampled from an interval $[x, x]$, i.e. if this interval was chosen, the value x would always be generated. This alternative solution could be explored in more depth in future research.

This thesis only considered scenarios where the original test scenario was carried out. The NPI methodology could be used in cases where both the original and the replicate experiment are carried out. Further study could investigate the potential conclusions to be drawn from the NPI reproducibility of both: the original and the replicate experiment, on the reproducibility of a second repeat of the experiment. The starting point of such exploration would be a practical application example.

Follow-up actions in case of low reproducibility are also important. Research into this has not yet been reported in the literature. The current definition of NPI reproducibility requires that the sample sizes of the original and the future samples are assumed to be equal. It would be of interest to study reproducibility with an adjusted definition, which would enable the future sample to be smaller or larger than the original sample, while all the other circumstances of the experiment would stay the same. This thesis has not addressed this, however, mathematically, this should not be a problem. The assumption of exchangeability might have to be rethought in such circumstances.

In conclusion, both the quantification of statistical reproducibility, and the use of the bootstrap method with small samples have potential to become more commonly used computational tools in preclinical research, helping decision-makers to make more informed choices. Given that translating preclinical to clinical research is a problem in pharmaceutical research, having more measures and tools available could improve efficiency in the field.

Appendix A

Additional material relevant to Chapter 2

A.1 The influence of using different quantile types upon the bootstrap performance in estimation

This is an extension to Section 2.4.2, where the performance in estimation of population characteristics was explored for the Normal distribution. The simulation study presented in Section 2.4.2 employed Type 7, $\hat{Q}_7(p)$, to calculate sample quantile and IQR, which is the default type in R. Hyndman and Fan [114] defined and compared different types of quantiles. Quantile types 4 through 9 can be used for continuous samples and this section briefly explores these.

For a distribution function, $F(x)$, quantile of a distribution is defined as follows [114]:

$$F^{-1}(p) = \inf\{x : F(x) \geq p\}, 0 < p < 1 \tag{A.1.1}$$

Based on an ordered sample of independent observations $X_{(1)}, \ldots, X_{(n_x)}$, sample quantiles provide estimation of their population counterparts [114]. Sample quantiles of type i can be written as [114]:

$$\hat{Q}_i(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)}, \text{where } \frac{j-m}{n_x} \leq p < \frac{j-m+1}{n_x}, 1 \leq i \leq 9 \tag{A.1.2}$$

for some $m \in \mathbb{R}$ and $0 \leq \gamma \leq 1$. Constant m is fixed for each sample quantile type. The value of γ is a function of $j = \lfloor pn_x + m \rfloor$ and $g = pn_x + m - j$. $\lfloor pn_x + m \rfloor$ stands for the

largest integer not less than $(pn_x + m)$ [114].

For types 4-9, $Q_i(p)$ is a continuous function of p, $\gamma = g$. The choice of p_k and m depends on the type i of quantile, $\hat{Q}_i(p)$.

$$\hat{Q}_4(p) : m = 0, p_k = \frac{k}{n_x} \tag{A.1.3}$$

$$\hat{Q}_5(p) : m = \frac{1}{2}, p_k = \frac{k - \frac{1}{2}}{n_x} \tag{A.1.4}$$

$$\hat{Q}_6(p) : m = p, p_k = \frac{k}{n_x + 1} \tag{A.1.5}$$

$$\hat{Q}_7(p) : m = 1 - p, p_k = (k-1)/(n_x - 1) \tag{A.1.6}$$

$$\hat{Q}_8(p) : m = \frac{(p+1)}{3}, p_k = \frac{k - \frac{1}{3}}{n_x + \frac{1}{3}} \tag{A.1.7}$$

$$\hat{Q}_9(p) : m = \frac{p}{4} + \frac{3}{8}, p_k = \frac{k - \frac{3}{8}}{n_x + \frac{1}{4}} \tag{A.1.8}$$

For small sample sizes, using different type of quantile seems to be making a difference on the calculation of the sample statistic. Figure A.1 displays Algorithm 1 outputs for the estimation of Q1 for $n = 4, 10$ for different quantile types. The biggest discrepancy is seen for Efron-B. For $n = 4$, the study conclusions remain the same, except that for $\hat{Q}_4(p)$ and $\hat{Q}_6(p)$, for $n = 4$, NPI-B is the best performing bootstrap method from the perspective of both metrics of assessment, whereas for the other quantile types, Banks-B has the lowest χ^2-value and NPI-B has the best coverage at 90% CI. For $n = 10$, the choice of the quantile type makes an impact on whether Banks-B or Hutson-B is the better performing bootstrap method. Thus, results vary when different types are used. Nevertheless, for most types, the actual recommendation of this thesis to use either Hutson-B or Banks-B for the estimation of median for Normally data remains. Similar conclusion is made for the estimation of Q3 for both $n = 4$ and $n = 10$. These considerations deserve further research attention.

The simulation outputs for the estimation of median for $n = 4, 10$ using different quantile types are presented in Figure A.2. From the figure, it can be inferred that the choice of the quantile type does not influence the conclusions regarding the bootstrap performance in the estimation of median for Normally distributed data: Banks-B is the best performing bootstrap method for $n = 4$, both Banks-B and Hutson-B perform similarly

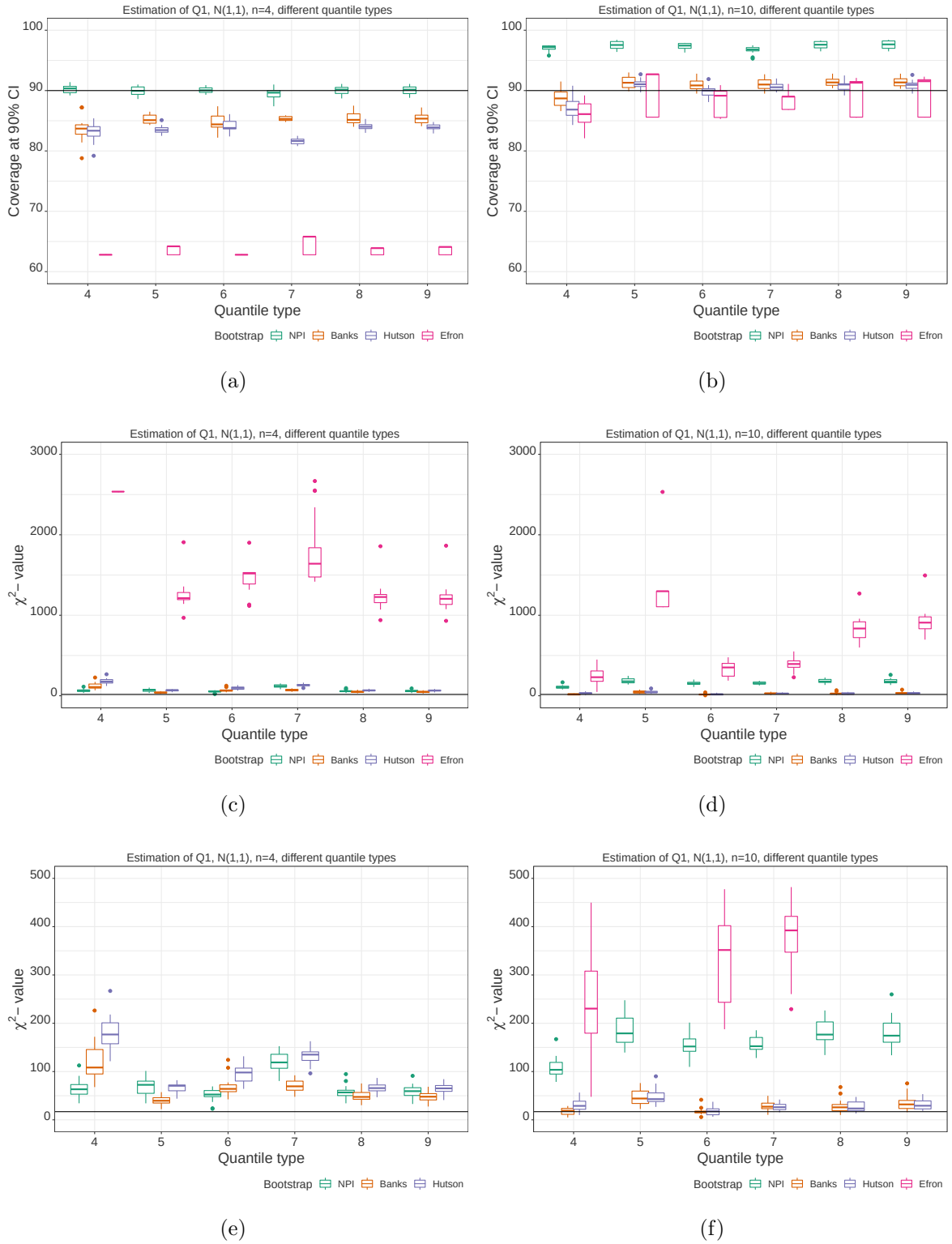


Figure A.1: Coverage at 90% CI and χ^2-values, estimation of Q1, N(1,1), $n = 4, 10$, finite (Approach I) NPI-B and Banks-B, different types of sample quantiles, 20 simulations

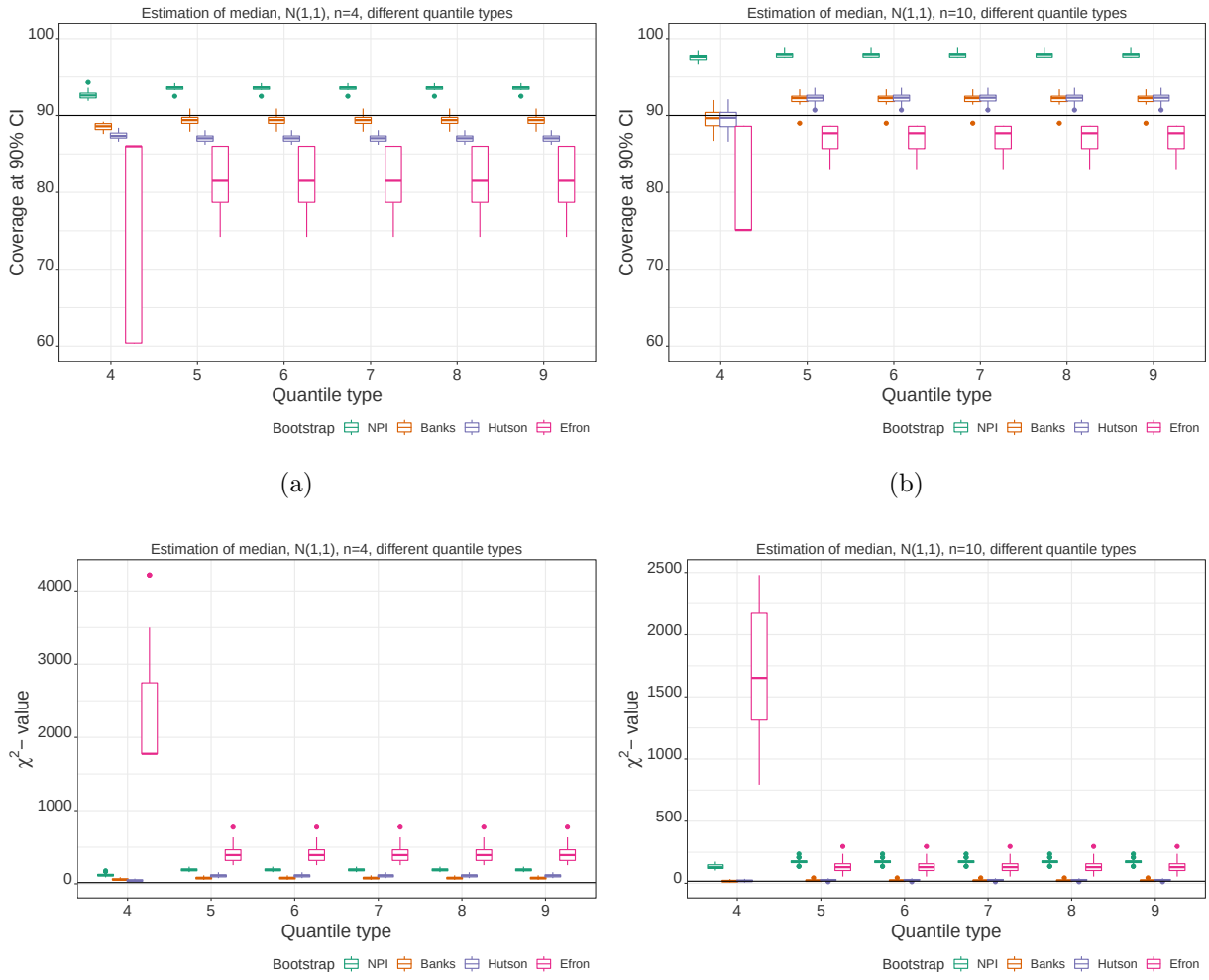


Figure A.2: Coverage at 90% CI and χ^2-values, estimation of median, N(1,1), $n = 4, 10$, finite (Approach I) NPI-B and Banks-B, different types of sample quantiles, 20 simulations

well for $n = 10$ and Efron-B is the worst performing bootstrap method. Efron-B performs particularly badly for $\hat{Q}_4(p)$. The Efron-B's performance in estimation, from the evaluation of both metrics of assessment, differs the most depending on what quantile type is used. Figure A.2 shows that in the estimation of median for $n = 4, 10$, using Type 4 makes Efron-B perform notably worse, from the perspective of both metrics of assessment. It is not a problem that Efron-B is the most affected bootstrap method as this work does not recommend the use of Efron-B for small samples anyway. The question of what type is the most suitable for calculating sample quantiles for small sample sizes remains outside the scope of this thesis. Further research could investigate this topic further.

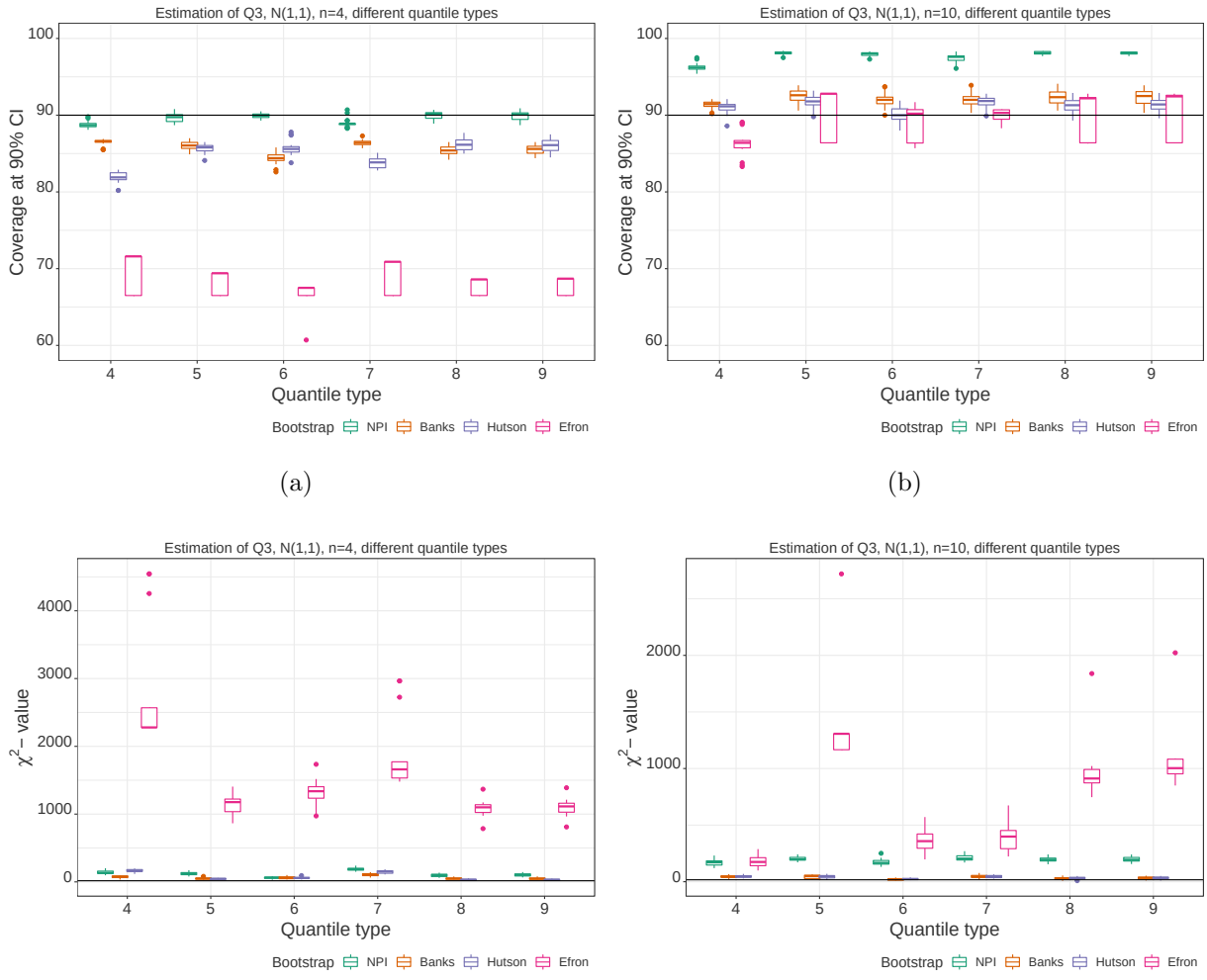


Figure A.3: Coverage at 90% CI and χ^2-values, estimation of Q3, N(1,1), $n = 4, 10$, finite (Approach I) NPI-B and Banks-B, different types of sample quantiles, 20 simulations

The choice of p_k has the biggest impact the performance in the estimation of IQR, especially on the χ^2-values and the coverage at 90% CI. For example, for $n = 4$, χ^2-value is the lowest for Banks-B for Type 7, for Banks-B for Types 4 and 5, and for Hutson-B for Types 6 and 8, as can be seen in Figure A.4. Given that this thesis does not recommend the use of small-sample bootstrap for the estimation of IQR, this discrepancy is not of a major concern.

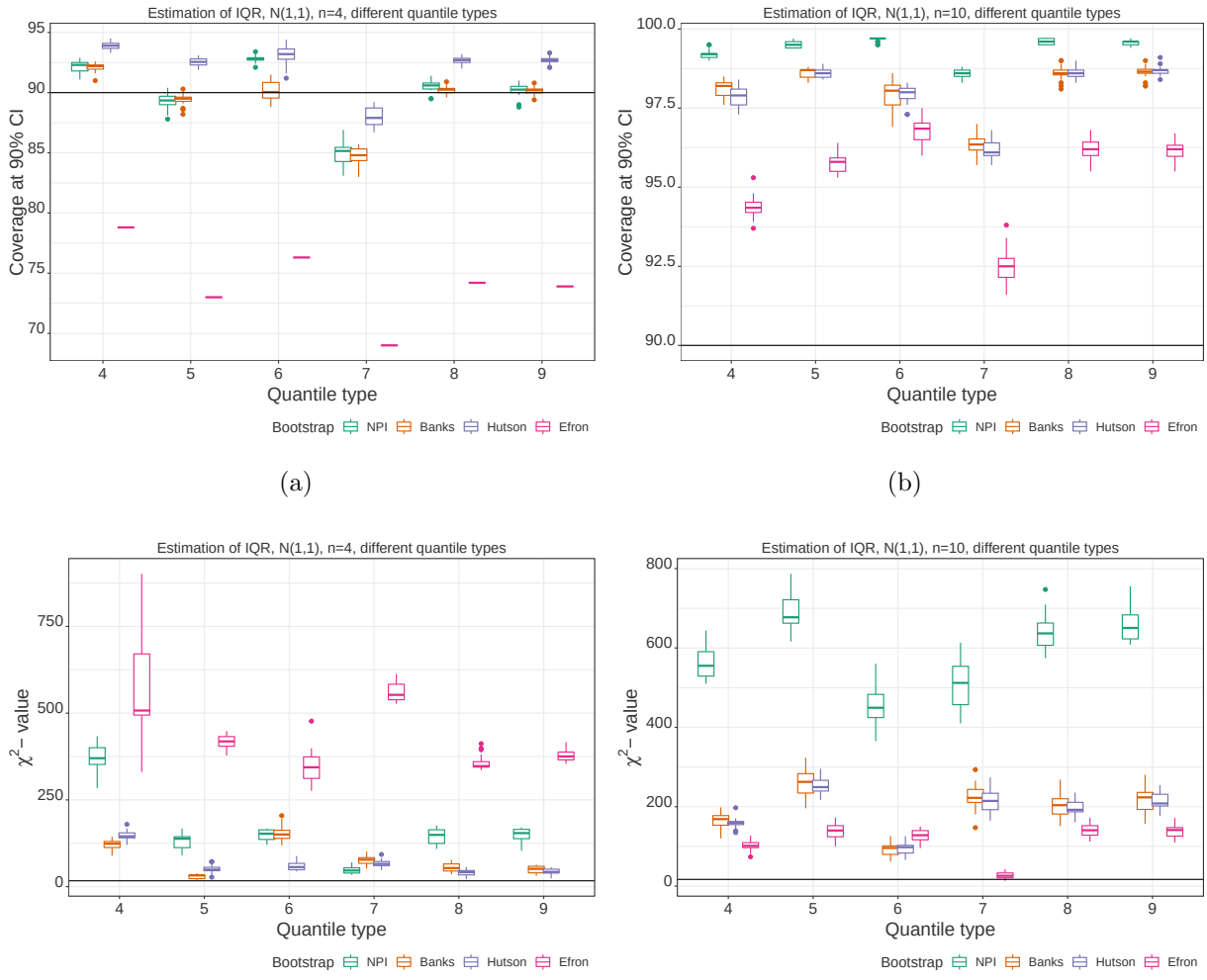
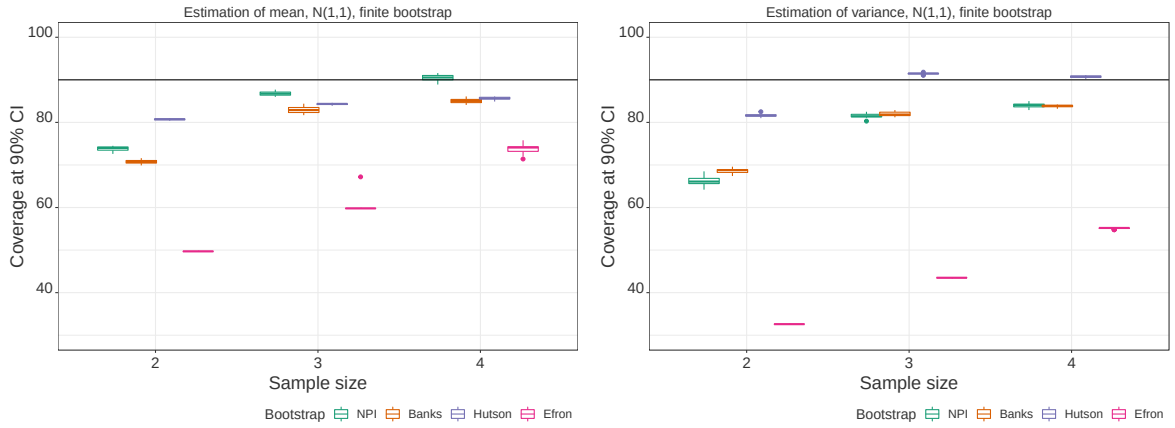


Figure A.4: Coverage at 90% CI and χ^2-values, estimation of IQR, N(1,1), $n = 4, 10$, finite (Approach I) NPI-B and Banks-B, different types of sample quantiles, 20 simulations

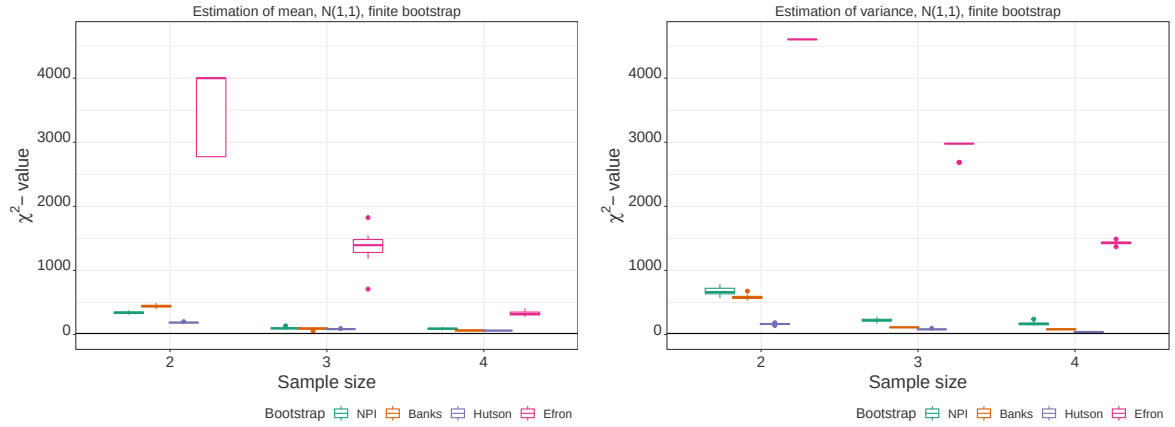
A.2 Bootstrap method performance for very small samples

A brief discussion regarding whether bootstrap methods can be used when the sample size is as small as $n = 3$, or even smaller, $n = 2$, follows. A simulation has been carried out to find out, using finite Approach I for Banks-B and NPI-B. The plots for simulation outcomes for sample sizes $n = 2, 3, 4$ are displayed in Figures A.5, A.6 and A.7, sample size $n = 4$ is included to show a pattern. As expected, Efron-B performs very poorly in the estimation of all the studied statistics for $n = 2, 3$ (very low coverage at 90% CI



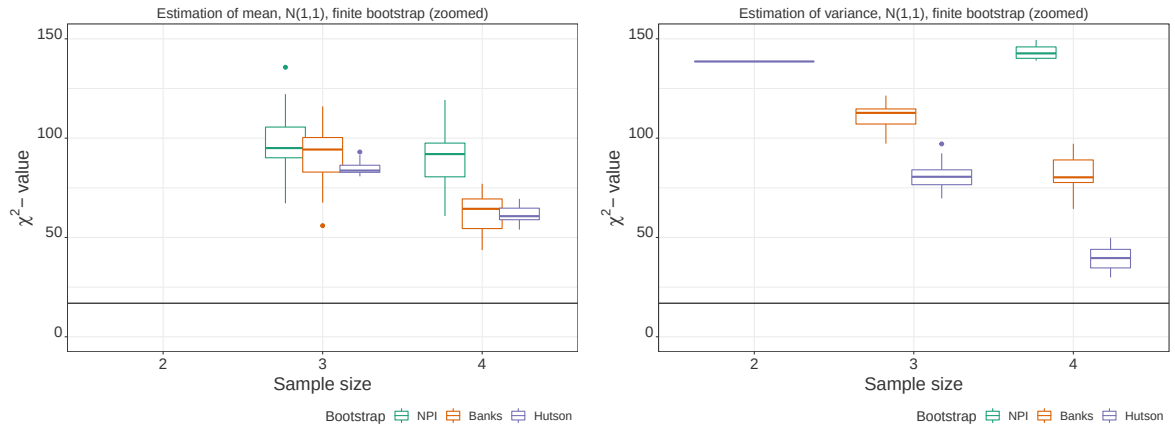
(a)

(b)



(c)

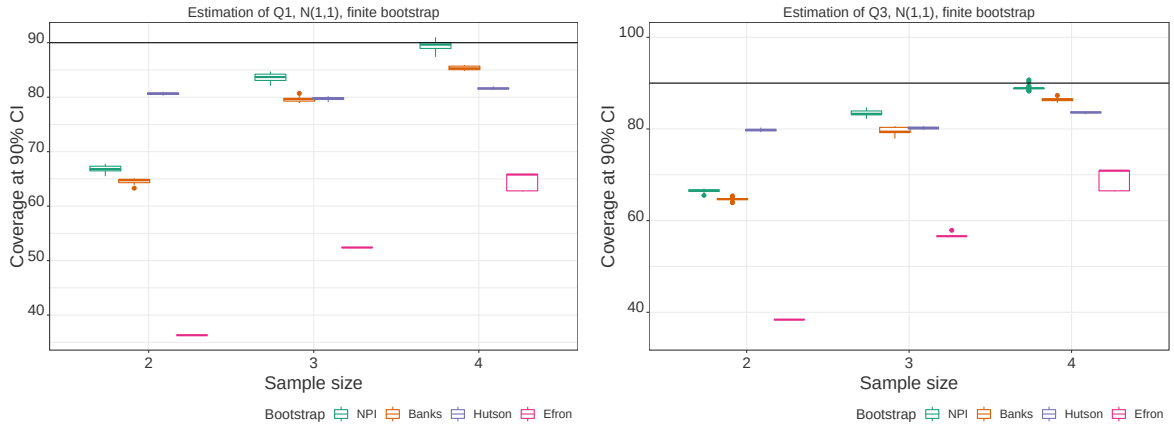
(d)



(e)

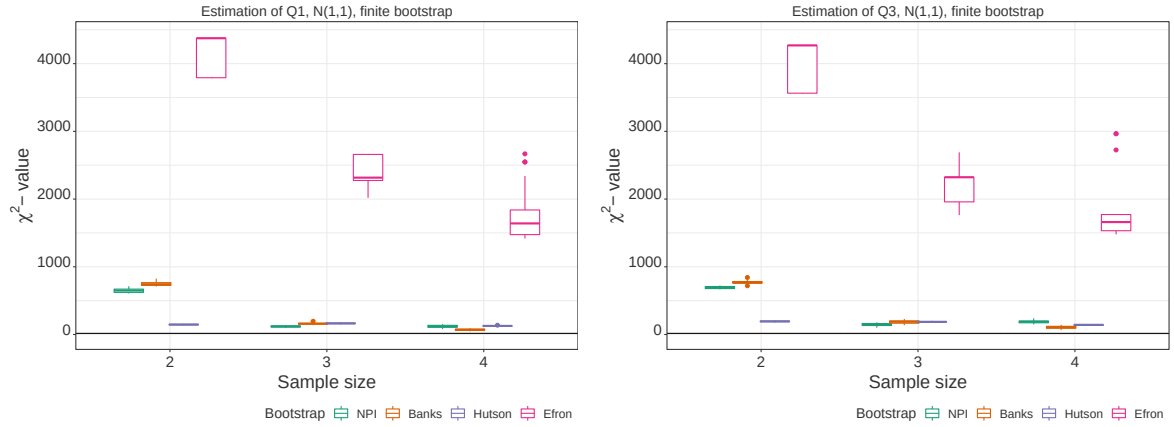
(f)

Figure A.5: Coverage at 90% CI and χ^2-values, estimation of mean and variance, N(1,1), $n = 2, 3, 4$, finite NPI-B and Banks-B, 20 simulations



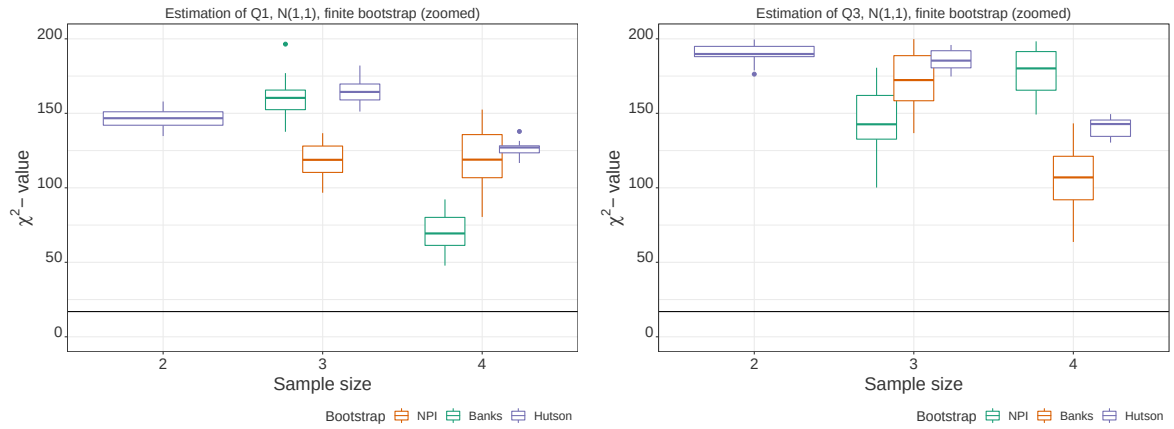
(a)

(b)



(c)

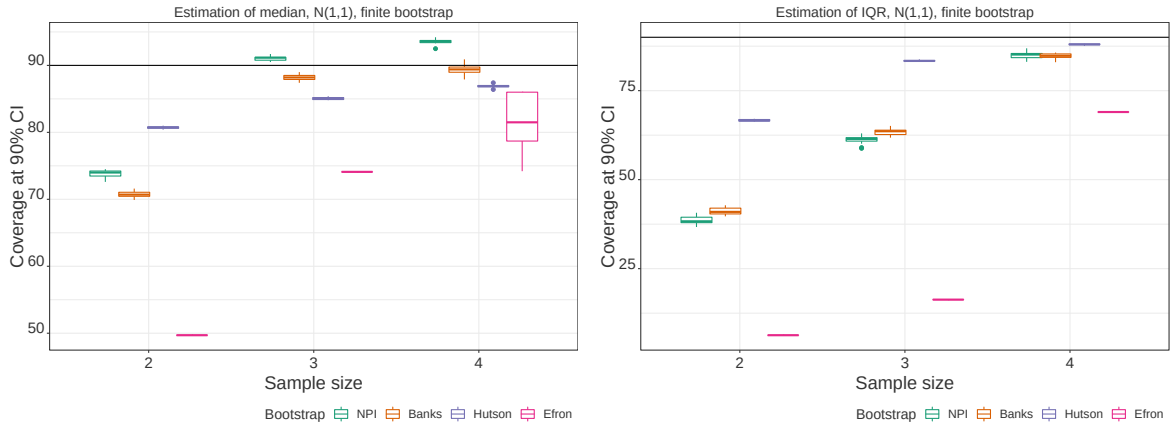
(d)



(e)

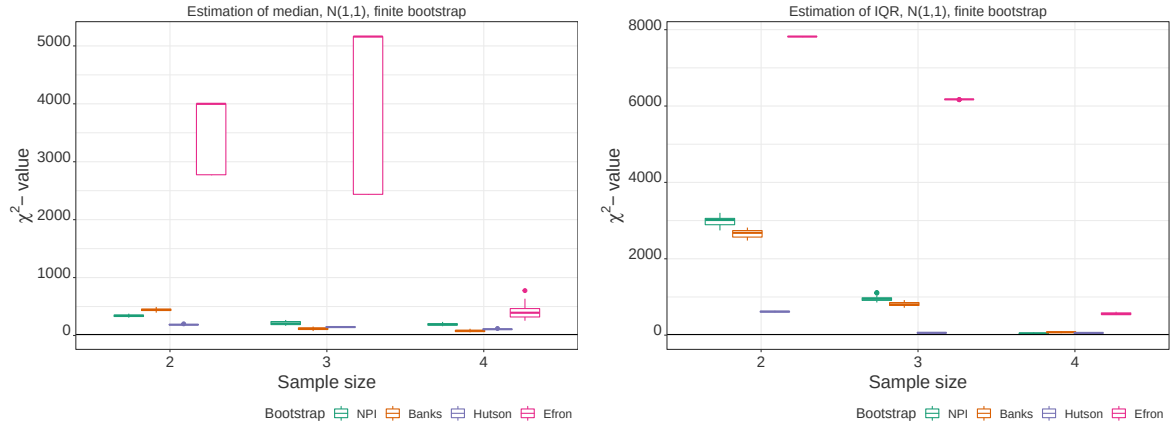
(f)

Figure A.6: Coverage at 90% CI and χ^2-values, estimation of Q1 and Q3, N(1,1), $n = 2, 3, 4$, finite NPI-B and Banks-B, 20 simulations



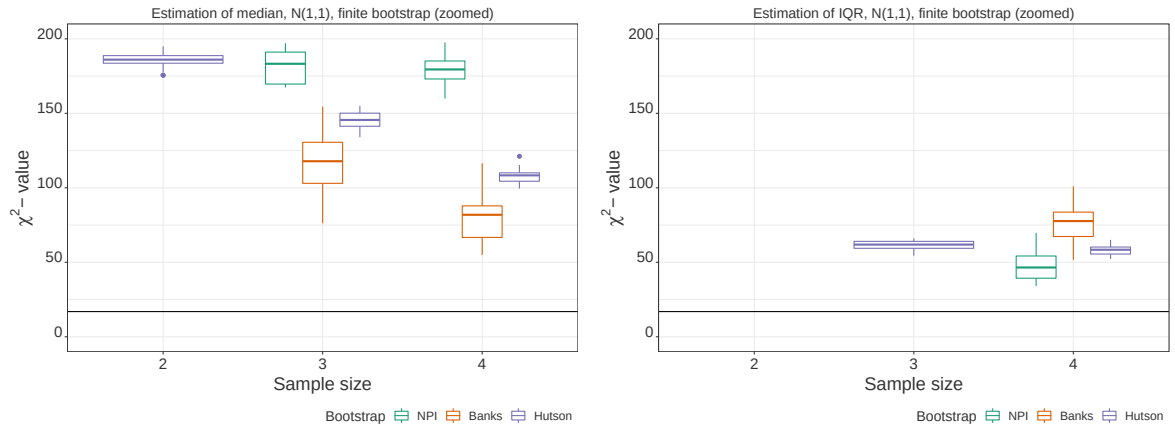
(a)

(b)



(c)

(d)



(e)

(f)

Figure A.7: Coverage at 90% CI and χ^2-values, estimation of median and IQR, N(1,1), $n = 2, 3, 4$, finite NPI-B and Banks-B, 20 simulations

and high χ^2-value). When the simulations are run for $n = 3$ for the estimation of mean (Figure A.5), there is under-coverage for all four bootstrap methods, however NPI-B has the lowest under-coverage at 90% CI and Hutson-B has the lowest χ^2-value, closely followed by Banks-B and NPI-B. Thus, Hutson-B and NPI-B perform relatively well in the estimation of mean for $n = 3$. For the estimation of mean for $n = 2$, Hutson-B is the best performing bootstrap, by considering both metrics of assessment. For the estimation of variance, for both sample sizes, $n = 2, 3$ (Figure A.5), Hutson-B is the best performing bootstrap, considering both metrics of assessment. At $n = 3$, Hutson-B has almost ideal coverage at 90% CI. In the estimation of Q1 and Q3 (Figure A.6), for $n = 3$, Banks-B is the best performing bootstrap, considering both metrics, but for $n = 2$, Hutson-B is the best performing bootstrap method. In the estimation of median (Figure A.7), Hutson-B is the best performing bootstrap method for $n = 2$ and for $n = 3$, NPI-B has the best coverage at 90% CI and Banks-B has the lowest χ^2-value, thus, it is unclear what is the best performing bootstrap method for $n = 3$ for the estimation of median. For the estimation of IQR (Figure A.7), Hutson-B is the best performing bootstrap method for both $n = 2, 3$. Overall, χ^2-values are still quite large for sample sizes $n = 2, 3$ for Banks-B, Hutson-B and NPI-B. It can be concluded that Banks-B, Hutson-B and NPI-B can be still considered for the use for the estimation of mean, variance and quantiles for Normally distributed data, however, the outcomes of such analysis should be considered with great care, given that there is no guarantee that such small sample is an accurate representation of the population. The bootstrap method should not be used for the estimation of IQR for $n = 3$ and lower.

A follow-up question is how the bootstrap method behaves when estimating mean for $n = 1$. This investigation has been done only out of theoretical interest, in practice bootstrap methods are not carried out for one observation. This topic is explored only for finite (Approach III) Banks-B and NPI-B because for those bootstrap methods a value is sampled from an intervals between observations, not from the sample observations. One datapoint is too small to estimate parameters required for finite (Approach I or II) and infinite NPI-B or Banks-B, or for Hutson-B. We created an interval around the one observation by subtracting and adding a value v to the one observation for $v = 0.1, 0.5, 1$ (finite Approach III, Section 2.3.3). The performance of NPI-B and Banks-B is very

Algorithm 11 Bootstrap variability in the estimation of statistics

1: Generate N datasets from a chosen distribution
2: For each dataset, generate B bootstrap samples;
3: Calculate the statistics for each of the bootstrap samples, $\hat{\theta}^*(b)$, for $b \in \{1, \ldots, B\}$;
4: Calculate the average bootstrapped statistics $\overline{\theta^*} = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$;
5: Calculate $s^2_{\hat{\theta}^*} = \sum_{b=1}^{B} (\hat{\theta}^*(b) - \overline{\theta^*})^2)/(B-1)$;
6: In total carry out Steps 2-5 N times, for all the generated datasets in Step 1.

similar for all v values. This is because for $n = 1$, Banks-B and NPI-B sample a value in exactly the same way (only from the second bootstrapped value, their algorithms differ). There is large under-coverage at 90% CI and large χ^2-value, but the performance in both metrics improves as v increases. For example, for $v = 0.5$, the average coverage at 90% CI is 34.5 % for both bootstrap methods and $\chi^2 = 3599$ for NPI-B and $\chi^2 = 3423$ for Banks-B.

A.3 Variability of bootstrap methods outcomes for Normally distributed data

This section reports the initial study of the variability of bootstrap methods outcomes. The variability is assessed via Algorithm 11. This algorithm is applied to each bootstrap method for mean, variance and median (Figure A.8) for small sample size ($n = 4, 6, 8, 10$). For mean and median, we study NPI-B and Banks-B with finite range (Approach I, Section 2.3.3) and for variance we study both finite range (Approach I, Section 2.3.3) and infinite range. N is set to 1000 and B is set to 1000. The conclusions for NPI-B, Banks-B and Efron-B samples are more consistent than for Hutson-B. Of these three bootstrap methods: NPI-B has the largest variance, Banks-B has the second largest variance and Efron-B has the smallest variance. Moreover, as n decreases, the variance of bootstrap outcomes decreases for all four bootstrap methods. BinHimd [31] explored this for three bootstrap methods (Efron-B, NPI-B and Banks-B) on finite range for sample sizes $n = 20, 50, 100, 200, 500, 1000$ and the conclusion for small samples are consistent with the conclusions for large samples. Choosing infinite range for the estimation of variance for

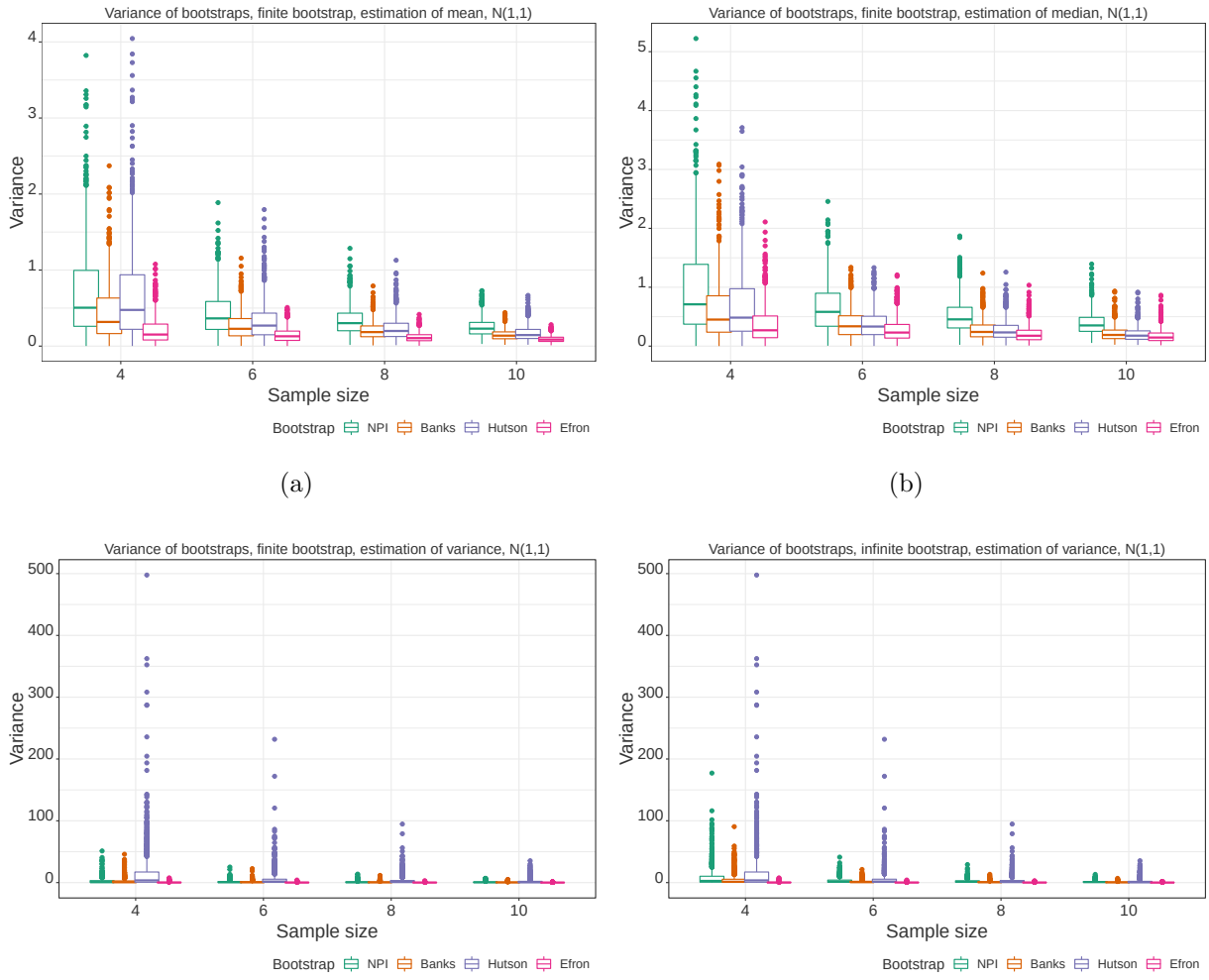


Figure A.8: Variability of bootstrap methods, estimation of mean, variance and variance, N(1,1), $n = 4, 6, 8, 10$

NPI-B and Banks-B leads to large variance of bootstrap samples. The difference is visible more for NPI-B and for Banks-B. The contrast between the finite and infinite range is the clearest for $n = 4$ out of all studied sample sizes.

Variance of Hutson-B outputs depends on the chosen population characteristic that is being estimated. In the estimation of mean, Hutson-B has larger variance than NPI-B for $n = 4$, similar variance to NPI-B for $n = 6$ and smaller variance than NPI-B for $n = 8, 10$, where it has larger variance than Banks-B and Efron-B. In the estimation of median, Hutson-B has the second largest variance (after NPI-B) for $n = 4$ and similar variance as Banks-B for $n = 6, 8, 10$. In the estimation of variance, Hutson-B has the largest variance

out of the four bootstrap methods. There might be a link between Hutson-B having the largest variance for the estimation of variance and being the best performing bootstrap method for the estimation of variance for Normally distributed data. Larger variance is typically not good for the estimation, as it may lead to over-coverage. However, this topic is left as a topic for future research.

A.4 The effect of the choice of range on the bootstrap method performance

A.4.1 Normally distributed data

Finite range (Approach I. and Approach III. with $v = 0.1$) and infinite range have been briefly discussed in Chapter 2. Figure A.9 also includes Approach II, which has been only introduced, but not used throughout Chapter 2. This figure illustrates that the choice of range makes an impact on the performance of both NPI-B and Banks-B in the estimation of mean for Normally distributed data. For different sample sizes, different range selection leads to the best bootstrap method performance. Finite Approach II is to a large extent influenced by the choice of c. For example, for the studied case, for both Banks-B and NPI-B, setting c to a large value, i.e. $c = 3$, leads to large χ^2-values and to over-coverage at 90% CI, this over-coverage increases as n increases. On the other hand, setting c to a very small value, i.e. $c = 0.2$, leads large under-coverage, especially for $n = 4, 6$. Further study into range choices for these two bootstrap methods is left as a topic for future research.

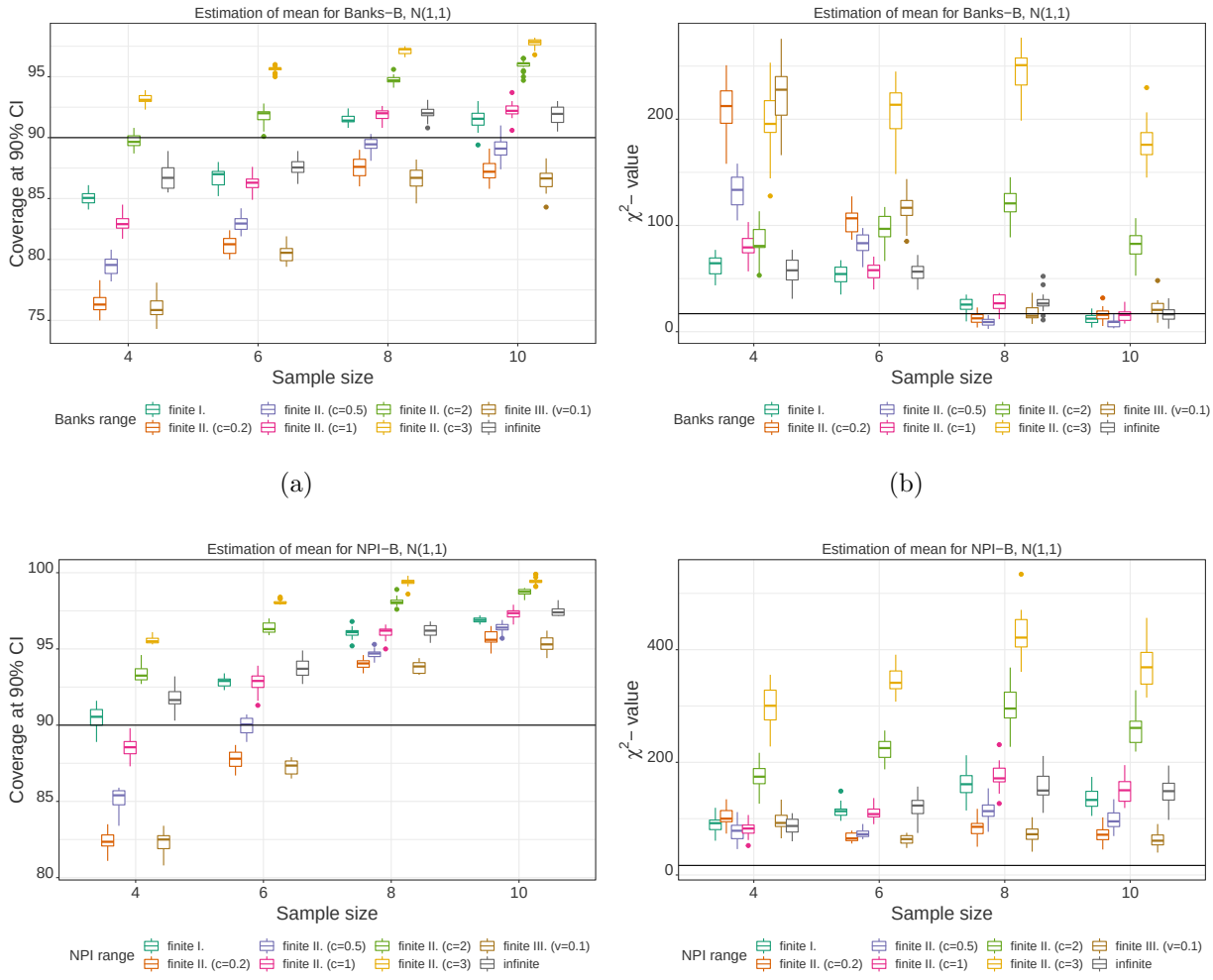


Figure A.9: Coverage at 90% CI and χ^2-values, estimation of mean, N(1,1), $n = 4, 6, 8, 10$, various ranges for NPI and Banks-B, 20 simulations

A.4.2 Exponentially and Lognormally distributed data

Section 2.4.3 used half-infinite range for data defined on $[0, \infty)$. The justification for this choice is provided here. From Figures A.10 and A.11, it can be inferred that Banks-B and NPI-B are better performing in estimation of variance of data from Exponential and Lognormal distribution when half-infinite range is used rather than when finite range is used. Thus, this work does not recommend the use of finite range for data defined on $[0, \infty)$ instead of half-infinite range.

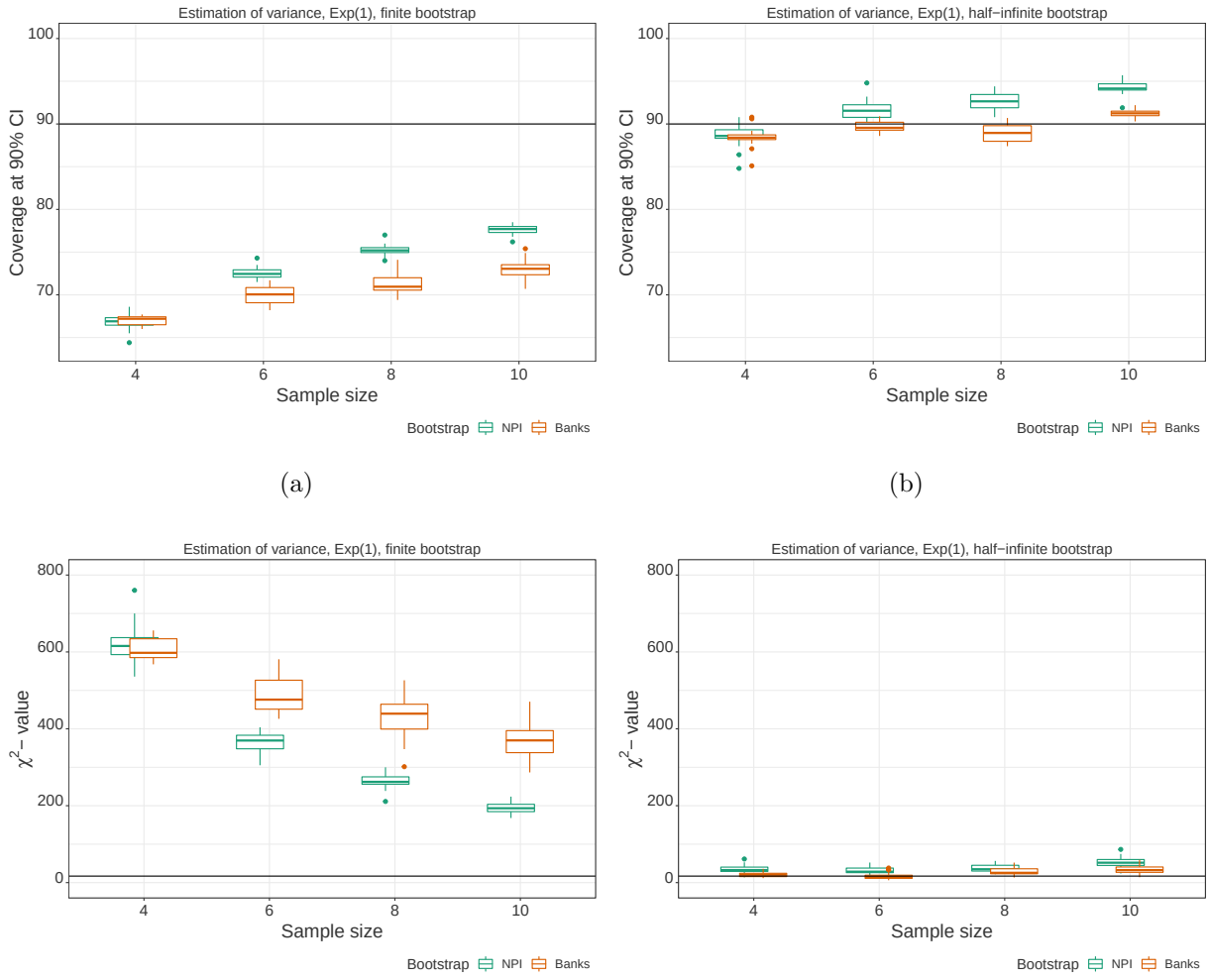


Figure A.10: Coverage at 90% CI and χ^2-values, estimation of variance, Exp(1), $n = 4, 6, 8, 10$, finite (Approach I) versus half-infinite NPI and Banks-B, 20 simulations

A.4.3 Mixed-Normally distributed data

In Section 2.4.4, finite range was assumed for Banks-B and NPI-B when estimating population characteristic for data from the Mixed-Normal distribution. Further simulations were carried out for Mixed-Normal A using infinite range for the estimation of mean (Figure A.12) and variance (Figure A.13). The conclusion of the additional simulation is that using infinite range does make a difference, especially for $n = 4$ where it improves coverage at 90% CI for both estimation of mean and variance. It also improves χ^2-values for all the studies sample sizes. The study of range for Mixed-Normal distribution could undergo more scrutiny, especially if a more complex simulation study is carried out in

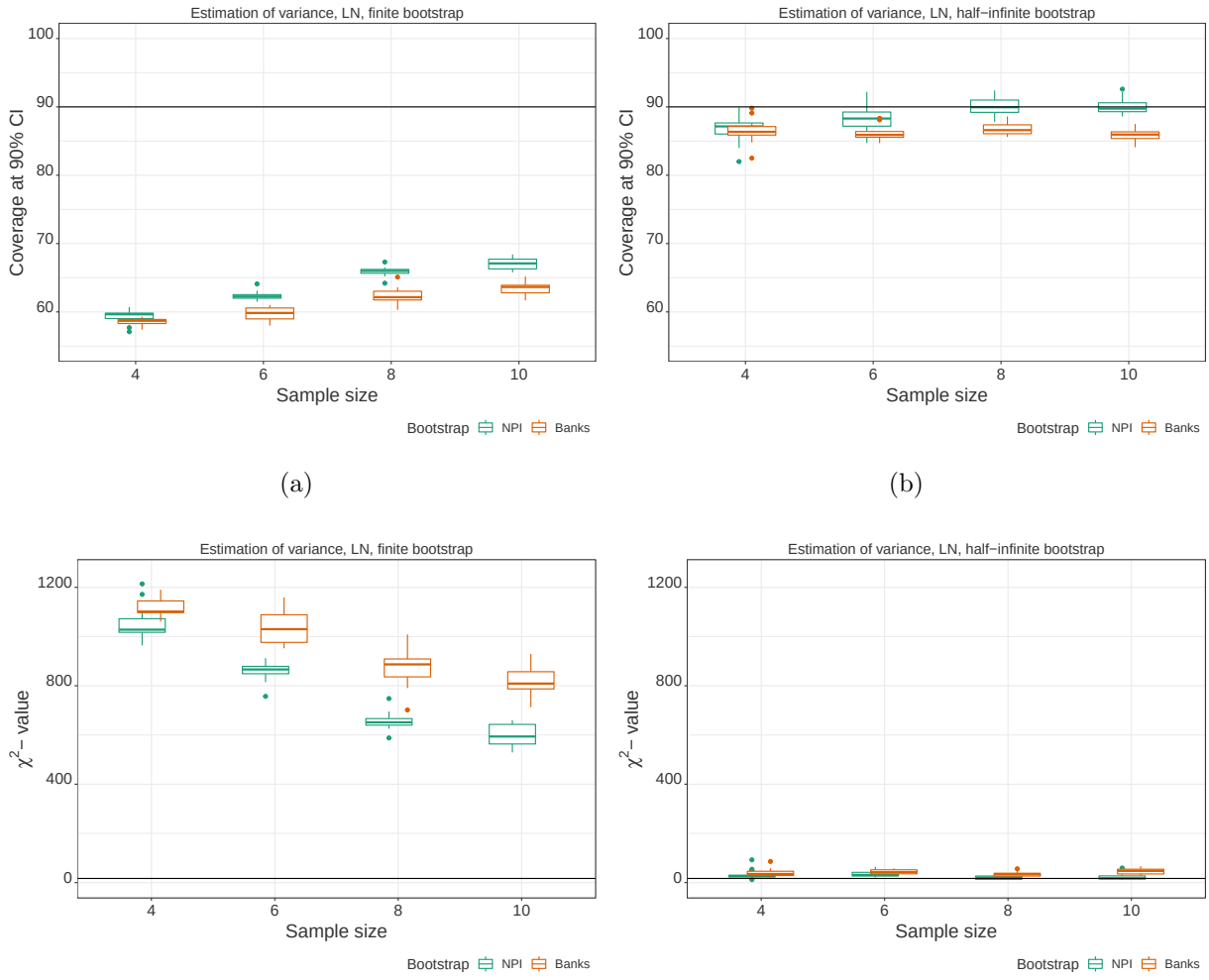


Figure A.11: Coverage at 90% CI and χ^2-values, estimation of variance, LN (m=1, sd=1), $n = 4, 6, 8, 10$, finite (Approach I) versus half-infinite NPI and Banks-B, 20 simulations

the future. Finite bootstrap is less computationally demanding, whereas infinite range improves coverage at $n = 4$ and reduces χ^2-values. If the reduction in χ^2-value is more important to the practitioner, than infinite range would be a reasonable choice.

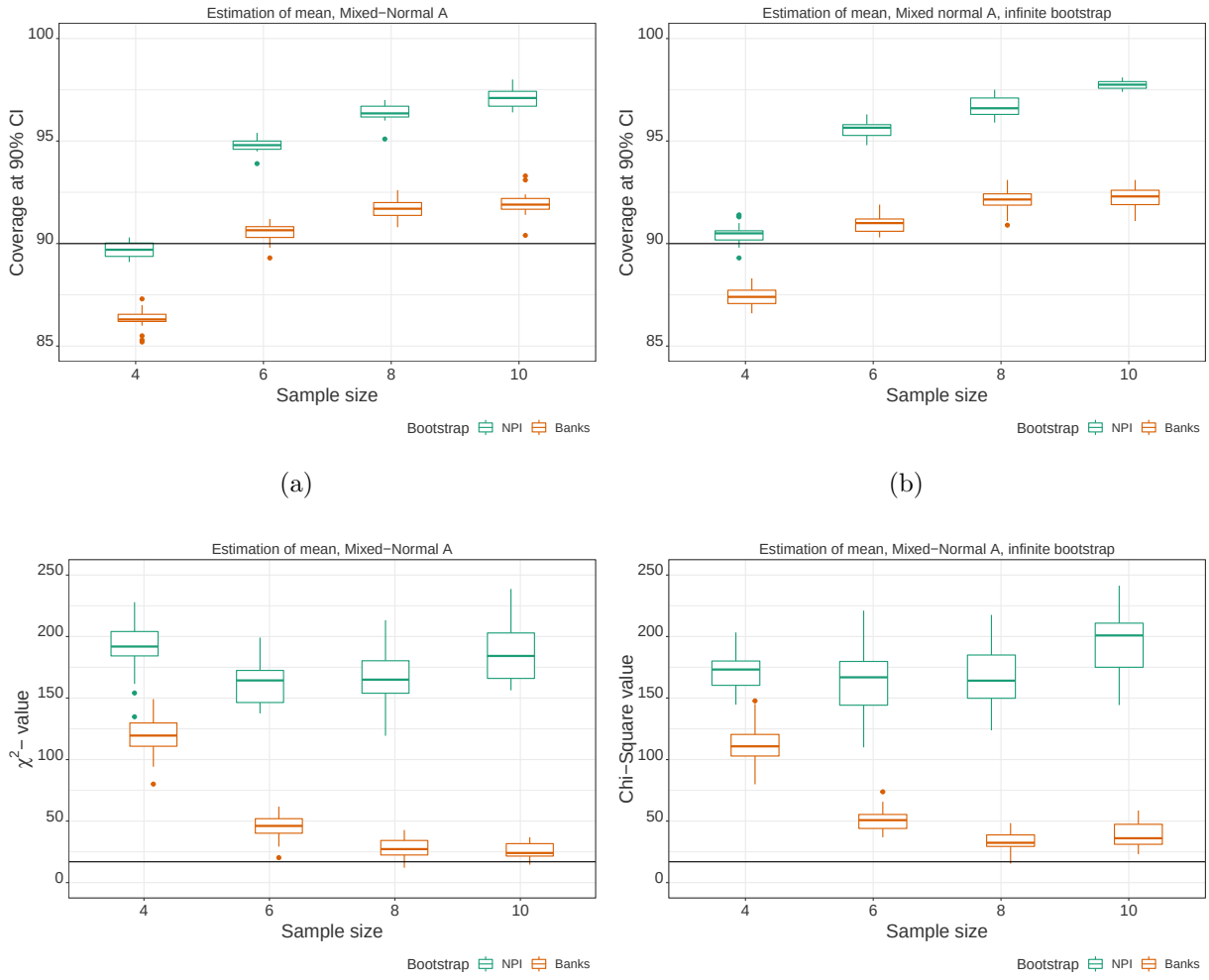


Figure A.12: Coverage at 90% CI and χ^2-values, estimation of mean, Mixed-Normal A, $n = 4, 6, 8, 10$, finite versus infinite bootstraps, 20 simulations

A.5 Performance in estimation of smoothed Efron-B by kernel (Kernel-B)

Chapter 2 focused on the bootstrap method performance of four bootstrap methods: Efron bootstrap, Banks bootstrap, NPI bootstrap and Hutson bootstrap. There is another bootstrap method that could be used with small sample sizes: smoothed Efron-B by kernel (Kernel-B). This bootstrap method was not included in the main investigation because its performance depends on the choice of smoothing parameter. However, this bootstrap method seems promising and, thus, some initial finding are reported here. Section A.5.1

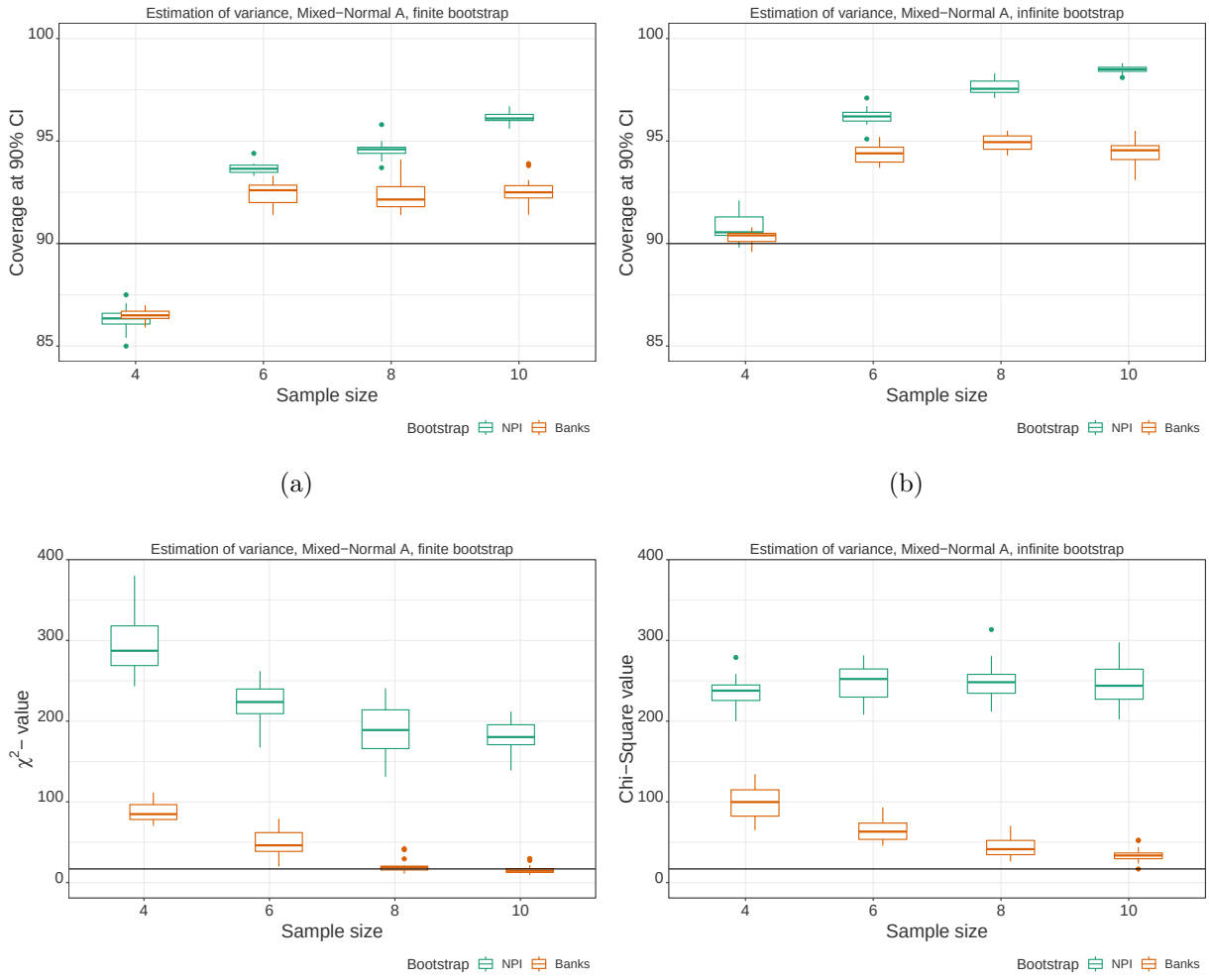


Figure A.13: Coverage at 90% CI and χ^2-values, estimation of variance, Mixed-Normal A, $n = 4, 6, 8, 10$, finite versus infinite bootstraps, 20 simulations

will introduce Kernel-B and Section A.5.2 will compare the performance of Kernel-B and Banks-B in the estimation of population characteristics for Normally, Exponentially and Lognormally distributed data.

A.5.1 Kernel-B

An alternative smoothened bootstrap method to Banks-B and Hutson-B is smoothened bootstrap using kernels (Kernel-B) [8, 76, 83, 168, 191], where the repeated resampling is performed from the smoothed version $\hat{F}$ of the empirical distribution of the observed data F_n. Kernel-B has received more attention than Banks-B and Hutson-B in the literature.

Polansky [167] explored the empirical coverage of Kernel-B, using bootstrap-t confidence intervals, for small samples ($n = 5, 10, 20$) for the mean, variance and correlation. Polansky [167] called this *smoothed bootstrap-t* and in the simulation study he compared it to Efron-B, using BC_a confidence intervals, and bootstrap-t confidence intervals, and additive corrected bootstrap-t. He considered 90% and 95% coverage for the two-sided interval. Polansky [167] concluded that for mean and variance, Kernel-B can decrease coverage error.

Silverman [190] defined univariate kernel density estimator as follows:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - X_i}{h}\Big) \tag{A.5.9}$$

where h is the bandwidth, x is a vector of data points used for estimating the kernel density and K is the kernel. There is a variety of kernel types to choose from, such as Epanechnikov, Biweight, Triangular, Gaussian and Rectangular [190]. This thesis only explores Gaussian kernel, see Equation (A.5.10), as the shape of this kernel does not change regardless of the bandwidth [62].

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, u \in \mathbb{R} \tag{A.5.10}$$

Kernel-B draws m data points with replacement from the initial dataset of size n, as with Efron-B. Then random noise from kernel density K is added to each of the drawn values. Repeating this procedure, in total B bootstrap samples are created. An estimate $\hat{\theta}$ of statistic θ is evaluated for each bootstrap sample. In more detail, the procedure to draw samples from univariate kernel density is as follows (Silverman [190]):

1. Calculate h using the original dataset.
2. Sample m data observations with replacement from n original data observations to create $x_1^*, x_2^*, \ldots, x_m^*$.
3. Generate m random values $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ from the kernel probability density K.
4. Create a bootstrap sample $y_1, y_2, \ldots, y_m$ by setting $y_i = x_i^* + h\varepsilon_i$ for $i = 1, \ldots, m$.
5. Repeat Steps 2 - 4 in total N times to create N Kernel-B samples.

Bandwidth h is also called window width or smoothing parameter in the literature. However, this thesis adopts the word bandwidth. The bandwidth determines the extent of smoothing. Larger bandwidth increases the variance proportionally. The choice of h in Silverman [190] is discussed with focus on h which minimises the approximate mean integrated square error. Note that Silverman's calculation of h was not created for smoothing the bootstrap, it was created for univariate density estimation. Wand and Jones [205] discussed methods for selecting the bandwidth: the plug-in bandwidth selection, the least squares cross-validation, the bias cross-validation, the estimation of density functionals and the smoothed cross-validation bandwidth selection. Similarly to the bandwidth selection discussed in Silverman [190], these are not specifically designed for the smoothed bootstrap method but for kernel smoothing in general. Kernel smoothing is a technique used for nonparametric estimation of functions [205]. This thesis limits the scope to the study of plug-in bandwidth selection and three different types of plug-in bandwidth selection are considered [190].

- Type 1: $h = 1.06\sigma_x n^{-\frac{1}{5}}$;
- Type 2: $h = 0.79IQR(x)n^{-\frac{1}{5}}$;
- Type 3: $h = \min(\sigma_x, 0.90(\frac{IQR(x)}{1.34})n^{-\frac{1}{5}})$.

The outputs of Kernel-B are affected by the choice of the kernel type and the bandwidth. This research does not focus on Kernel-B because the outputs of bootstrap methods are affected by these choices and also by the underlying distribution of the data. As already discussed, for small sample sizes it is not possible to ascertain the exact underlying distribution of the data. Nevertheless, Section A.5.2 briefly compares Banks-B and Kernel-B. The performance of Kernel-B in the estimation of mean, median, and variance for samples $n \leq 10$ is considered. Comparison of these two smoothed bootstrap method has not been explored in the literature yet. For Kernel-B implementation, the `kernelboot` R package can be used, however, this function has limited choice of the smoothing parameter for univariate one-dimensional samples. Thus, a new function was written in R.

A.5.2 Kernel-B vs. Banks-B

A small simulation study was run to explore the performance of Kernel-B in estimation. Three different ways of calculating the plug-in estimate of bandwidth, defined in Section A.5.1, were explored. Kernel-B with some types of bandwidth selection performed well in estimation, especially in the estimation of median. The simulation outputs of Kernel-B alongside outputs for Banks-B in the estimation of mean, variance, Q1, median, Q3 and IQR for Normally, Lognormally and Exponentially distributed data are displayed in Figures A.14, A.15, A.16 A.17, A.18 and A.19, respectively.

This initial study concluded that Kernel-B with different types of bandwidth performed differently in the estimation of different population characteristics, for different sample sizes and for data with different underlying distributions. Table A.1 sums up what bootstrap methods, a choice from Banks-B and Kernel-B Type 1 to Type 3, performed the best at various circumstances (different sample sizes, different population characteristics of interest, different underlying distributions). For Exponentially and Lognormally distributed data, for all sample sizes, Banks-B performs better than all types of Kernel-B in the estimation of mean and variance. Another inference made from the figures is that Kernel-B performs better for the Normal distribution than it does for the Exponential or Lognormal distribution. Even though there are cases where Kernel-B performs better than Banks-B, for example when estimating median at n for all three studied distributions, there is not one type of Kernel-B which would perform consistently well in the estimation of one particular population characteristic across different sample sizes.

Given that for small sample sizes Kernel-B does not perform well in the estimation of mean and variance for some of the studied distributions (i.e. Exponential and Lognormal), this thesis would not recommend the use of Kernel-B for the estimation of mean or variance. This study did not provide clear recommendations on Kernel-B use for the estimation of quantiles and IQR. Further research into the performance of Kernel-B in the estimation of quantiles (Q1, median, and Q3) would be meaningful. Moreover, further study could investigate the use of other bandwidths for Kernel-B when making inference for small samples.

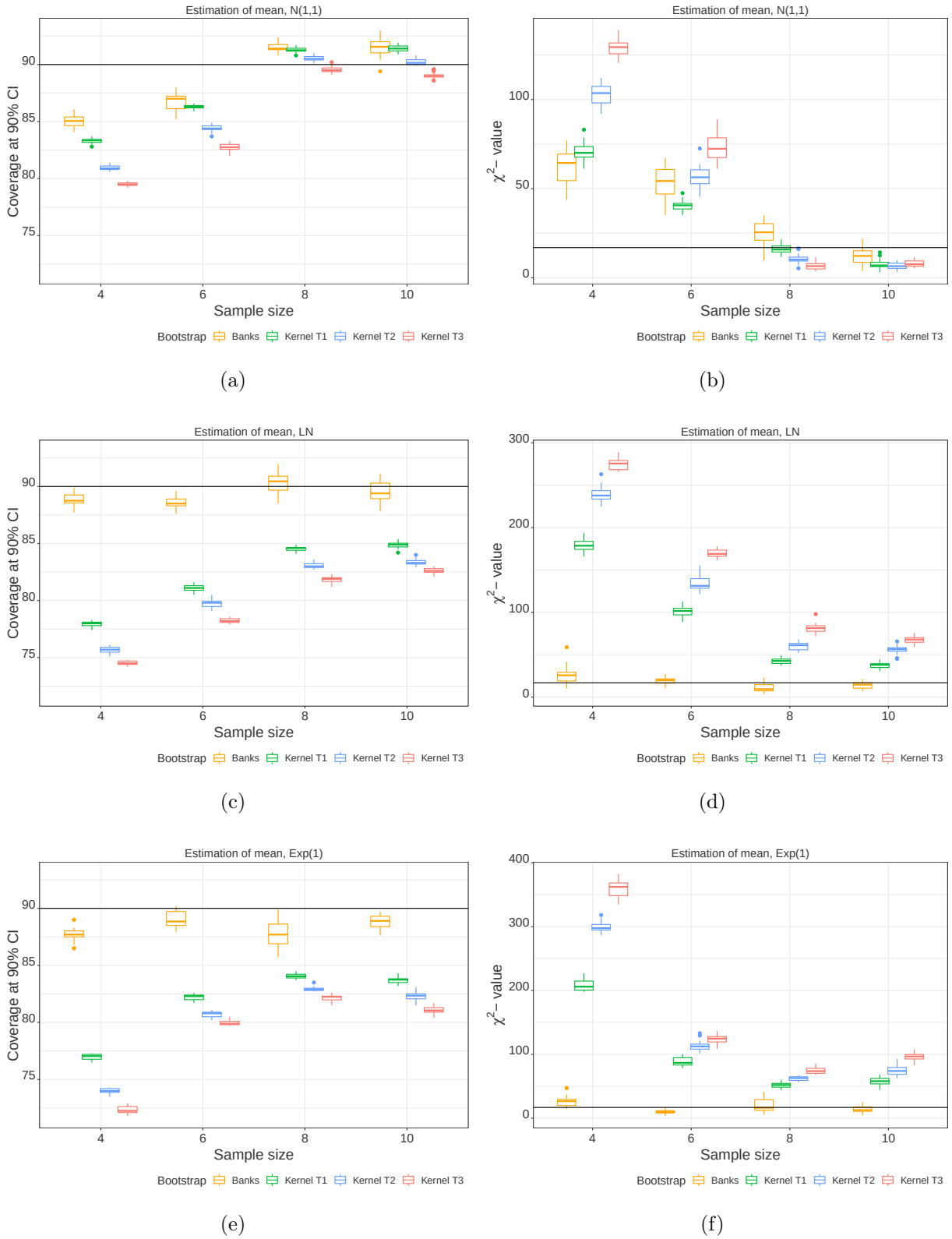


Figure A.14: Coverage at 90% CI and χ^2-values, estimation of mean, N(1,1), LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, Banks-B vs. Kernel-B (T1, T2, T3), 20 simulations

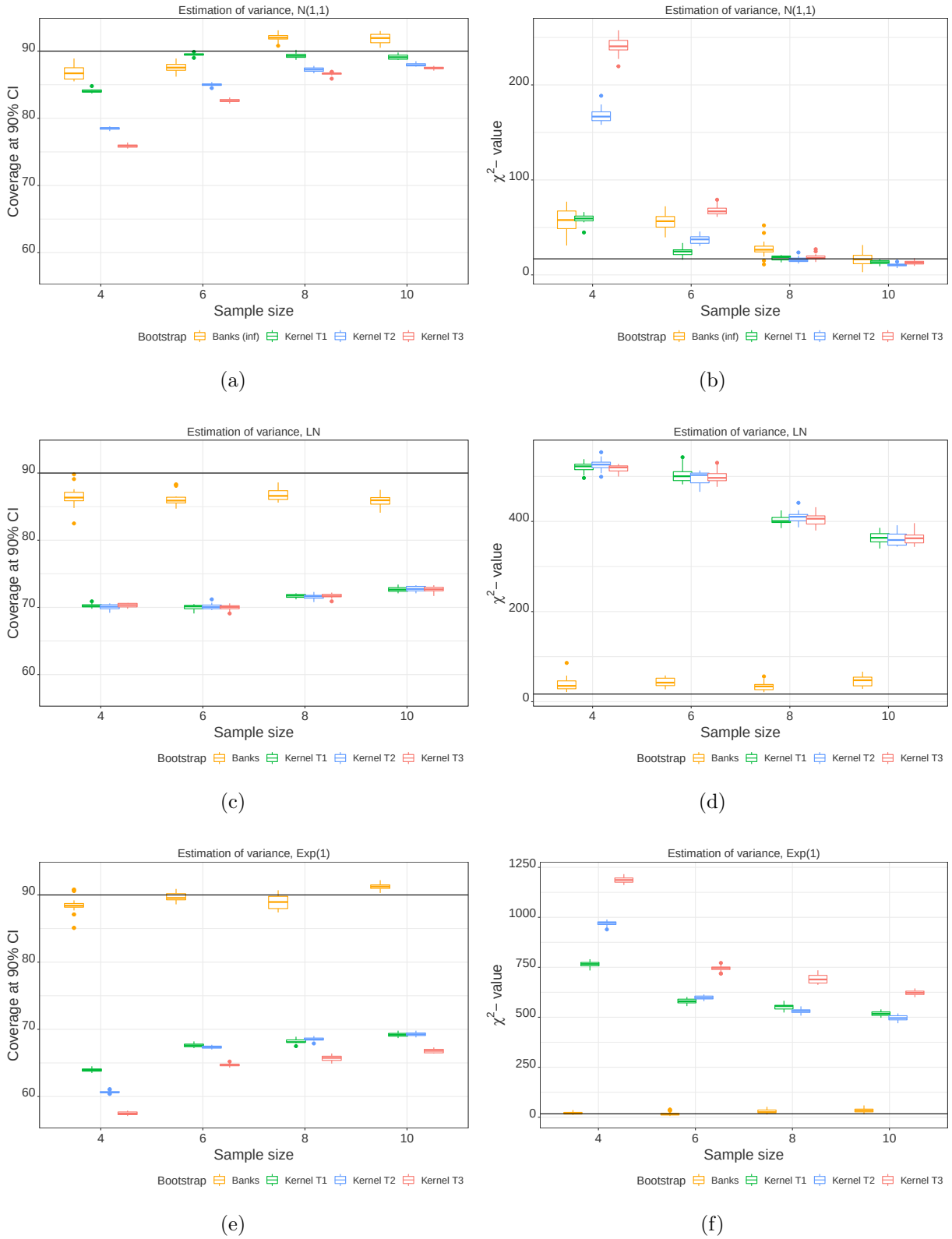


Figure A.15: Coverage at 90% CI and χ^2-values, estimation of variance, N(1,1), LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1) , $n = 4, 6, 8, 10$, Banks-B vs. Kernel-B (T1, T2, T3), 20 simulations

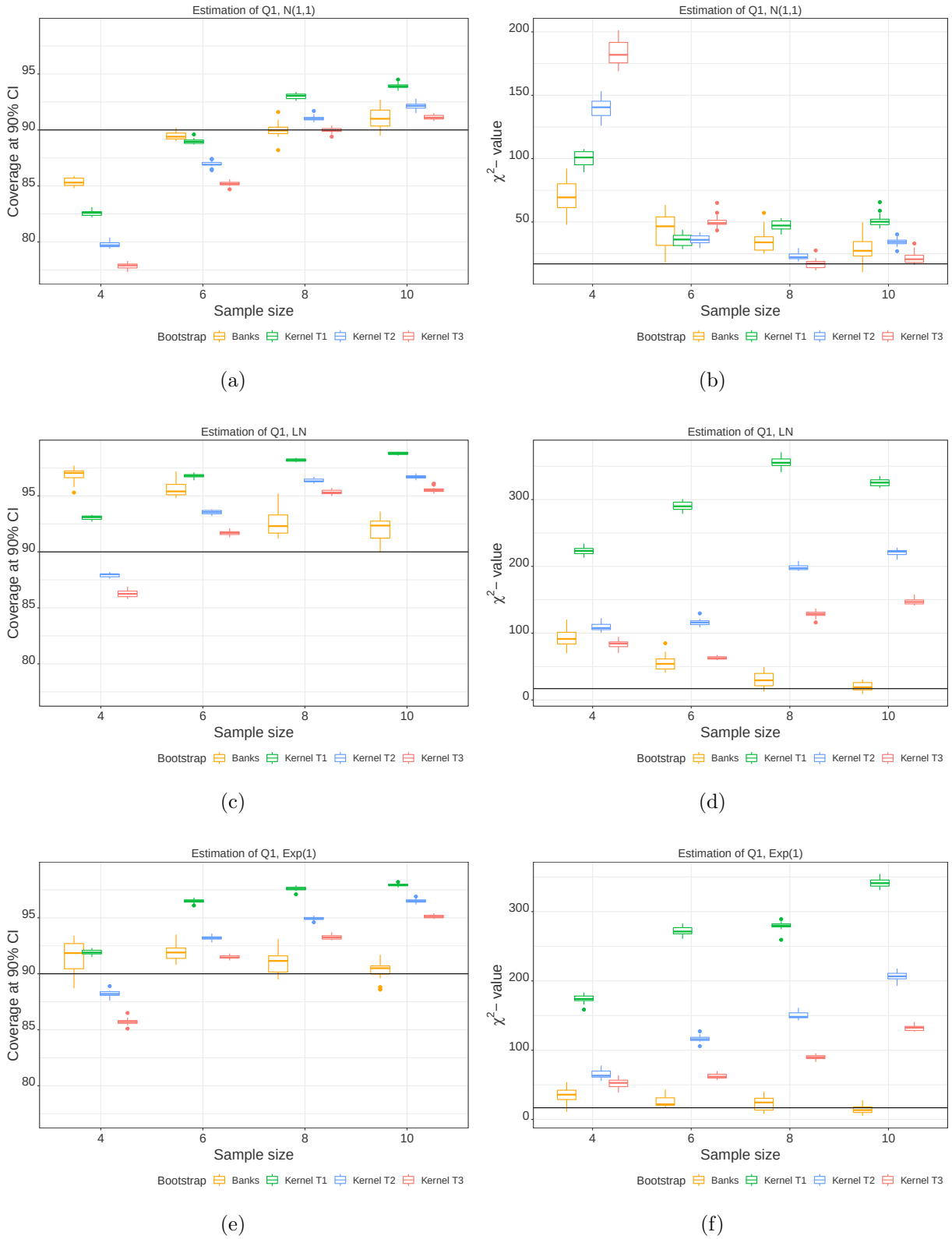


Figure A.16: Coverage at 90% CI and χ^2-values, estimation of Q1, N(1,1), LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, Banks-B vs. Kernel-B (T1, T2, T3), 20 simulations

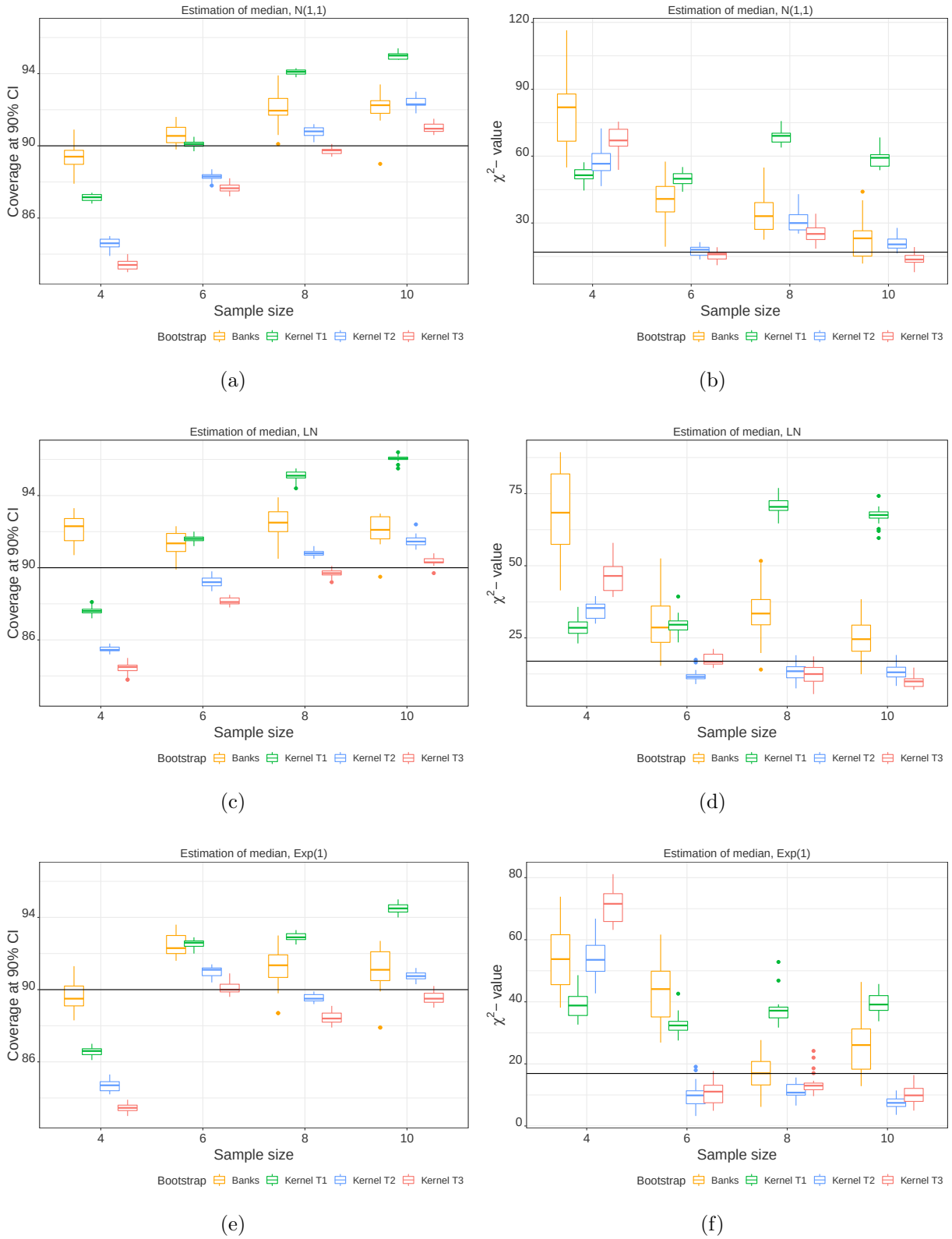


Figure A.17: Coverage at 90% CI and χ^2-values, estimation of median, N(1,1), LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, Banks-B vs. Kernel-B (T1, T2, T3), 20 simulations

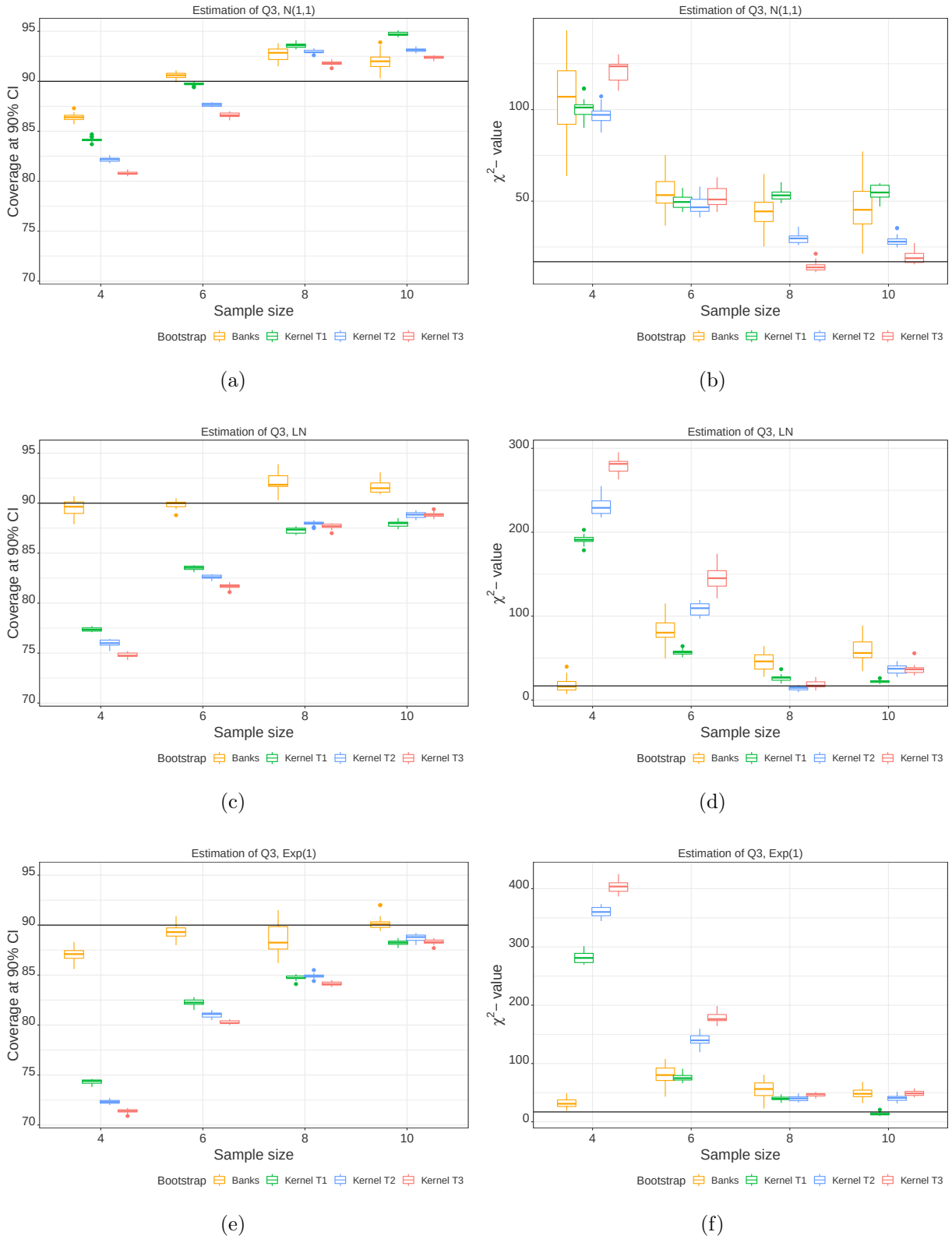


Figure A.18: Coverage at 90% CI and χ^2-values, estimation of Q3, N(1,1), LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, Banks-B vs. Kernel-B (T1, T2, T3), 20 simulations

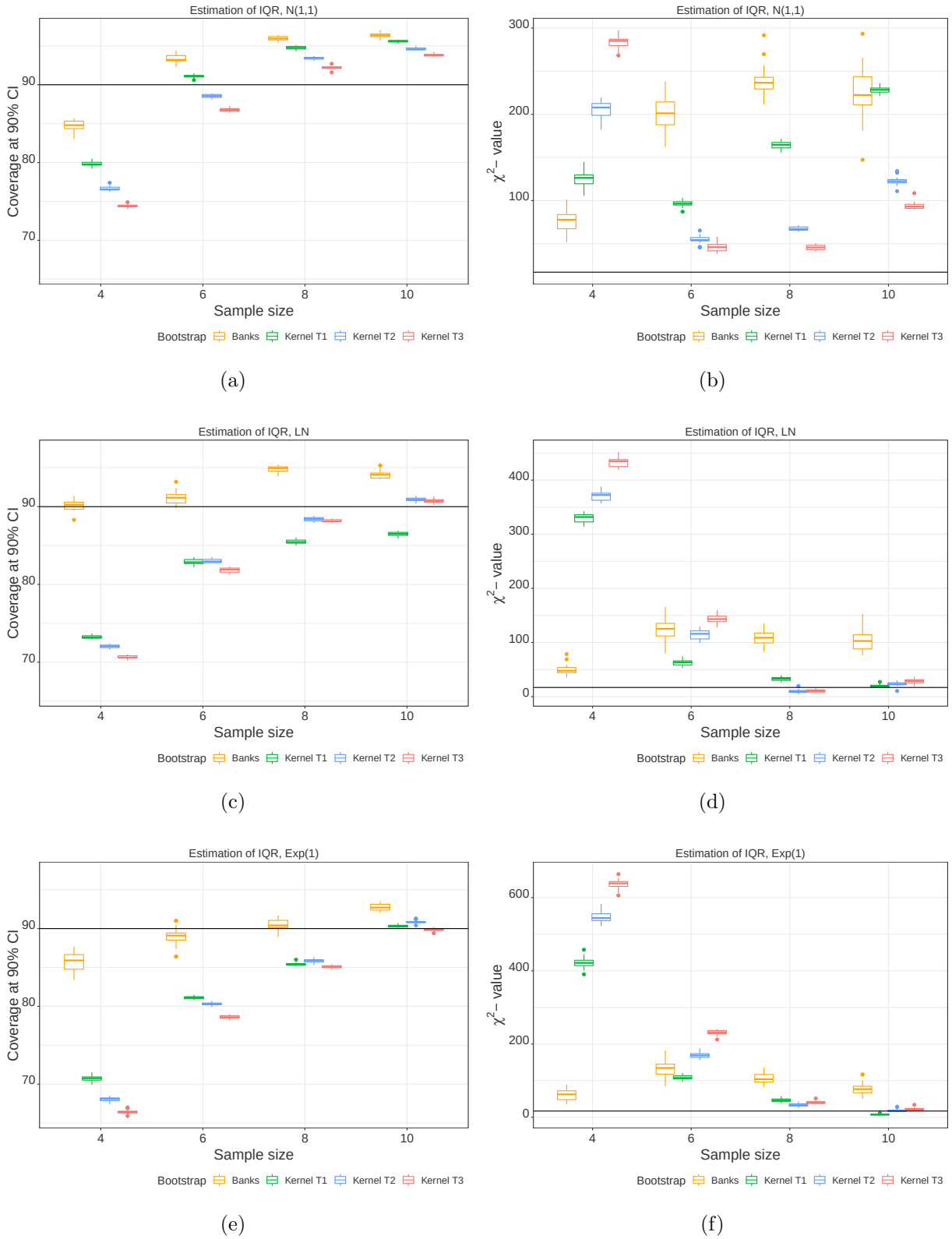


Figure A.19: Coverage at 90% CI and χ^2-values, estimation of IQR, N(1,1), LN($m_{LN} = -0.347$, $s^2_{LN} = 0.833^2$) and Exp(1), $n = 4, 6, 8, 10$, Banks-B vs. Kernel-B (T1, T2, T3), 20 simulations

Estimated characteristic	Sample size	Distribution		
		Normal	**Lognormal**	**Exponential**
Mean	4	Banks-B	Banks-B	Banks-B
	6	Banks-B	Banks-B	Banks-B
	8	Kernel-B-T2	Banks-B	Banks-B
	10	Kernel-B-T2	Banks-B	Banks-B
Variance	4	Banks-B	Banks-B	Banks-B
	6	Kernel-B-T1	Banks-B	Banks-B
	8	Banks-B	Banks-B	Banks-B
	10	Banks-B	Banks-B	Banks-B
Q1	4	Banks-B	Banks-B	Banks-B
	6	Kernel-B-T1	Kernel-B-T3	Banks-B
	8	Kernel-B-T3	Banks-B	Banks-B
	10	Kernel-B-T3	Banks-B	Banks-B
Median	4	Banks-B	Banks-B	Banks-B
	6	Kernel-B-T1	Banks-B	Banks-B
	8	Kernel-B-T2	Banks-B	Kernel-B-T2
	10	Kernel-B-T3	Kernel-B-T3	Kernel-B-T3
Q3	4	Banks-B	Banks-B	Banks-B
	6	Banks-B	Banks-B	Banks-B
	8	Kernel-B-T3	Banks-B	Banks-B
	10	Kernel-B-T3	Banks-B	Banks-B
IQR	4	Banks-B	Banks-B	Banks-B
	6	Kernel-B-T1	Banks-B	Banks-B
	8	Kernel-B-T3	Kernel-B-T2	Banks-B
	10	Kernel-B-T3	Kernel-B-T2	Kernel-B-T1

Table A.1: Table summarising whether the Banks-B or Kernel-B method performs the best in the estimation of various population characteristics for various small sample sizes ($n = 4, 6, 8, 10$) and for various distributions (Normal, Lognormal, Exponential)

Appendix B

Additional material relevant to Chapter 4

B.1 Reproducibility for the pairwise t-test

This section provides additional Tables to Section 4.3. NPI-B-RP for Approach II (with $c = 2$) is presented in Table B.1. These outputs were displayed in Section 4.3 in the form of a plot. The values of finite NPI-B (Approach I) and $2 * IQR$ (Approach II, $c = 2$) for each dose are presented in Table B.2. Table B.3 shows NPI-B-RP outputs for Approach II (with $c = 1.5, 3$), Tables B.4 and B.5 display NPI-B-RP for t-test and WMT, respectively, and Table B.6 shows NPI-B-RP for the infinite approach. The p-values in Tables B.1 and B.6 are for t-test, without BH adjustment.

Pairwise	Statistics of the real data				Algorithm 1 output					
	Reject	p-value	ES	Cohen's	t-test			WMT		
	?			d	min	mean	max	min	mean	max
A vs. B	Yes	0.0003	0.226	2.041	0.831	0.856	0.883	0.822	0.850	0.881
B vs. C	Yes	0.0000	0.366	3.213	0.963	0.976	0.989	0.941	0.959	0.969
C vs. D	Yes	0.0007	0.178	1.778	0.680	0.718	0.755	0.724	0.762	0.789
D vs. E	Yes	0.0191	0.097	1.038	0.455	0.497	0.532	0.514	0.553	0.581
E vs. F	No	0.5977	-0.013	-0.115	0.883	0.913	0.935	0.918	0.936	0.952

Table B.1: Statistical and reproducibility analysis for all test scenario's pairwise comparisons - Approach II (with $c = 2$)

Dose	Maxium distance	2*IQR
A	0.193	0.149
B	0.071	0.376
C	0.085	0.393
D	0.105	0.112
E	0.098	0.226
F	0.143	0.186

Table B.2: Maximal distance and 2*IQR for each dose used in finite bootstrap

Pairwise	Algorithm 1 output											
	t-test						WMT					
	1.5*IQR			3*IQR			1.5*IQR			3*IQR		
	min	mean	max	min	mean	max	min	mean	max	min	mean	max
A vs. B	0.871	0.895	0.913	0.747	0.774	0.804	0.852	0.874	0.897	0.775	0.807	0.837
B vs. C	0.984	0.992	0.997	0.898	0.924	0.941	0.972	0.984	0.994	0.899	0.920	0.946
C vs D	0.740	0.772	0.807	0.590	0.631	0.664	0.757	0.785	0.811	0.705	0.740	0.777
D vs E	0.497	0.536	0.595	0.552	0.586	0.622	0.532	0.566	0.601	0.505	0.534	0.574
E vs F	0.888	0.913	0.934	0.894	0.912	0.940	0.916	0.935	0.953	0.915	0.935	0.955

Table B.3: Reproducibility analysis for all test scenario's pairwise comparisons - Approach II (with $c = 1.5$ and $c = 3$)

Pairwise	v=0.1			v=0.05			v=0.01		
	min	mean	max	min	mean	max	min	mean	max
A vs. B	0.939	0.956	0.970	0.964	0.977	0.987	0.975	0.988	0.996
B vs. C	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
C vs. D	0.836	0.876	0.906	0.890	0.911	0.928	0.918	0.937	0.957
D vs. E	0.550	0.589	0.619	0.587	0.627	0.662	0.625	0.657	0.707
E vs. F	0.891	0.912	0.929	0.889	0.912	0.931	0.878	0.911	0.932

Table B.4: NPI-B-RP for the t-test, all test scenario's pairwise comparisons - Approach III

Pairwise	v=0.1			v=0.05			v=0.01		
	min	mean	max	min	mean	max	min	mean	max
A vs. B	0.885	0.912	0.944	0.917	0.940	0.954	0.931	0.953	0.971
B vs. C	0.998	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000
C vs. D	0.830	0.854	0.879	0.853	0.882	0.904	0.870	0.897	0.924
D vs. E	0.568	0.610	0.645	0.612	0.651	0.682	0.645	0.682	0.740
E vs. F	0.916	0.933	0.952	0.909	0.932	0.949	0.913	0.935	0.960

Table B.5: NPI-B-RP for WMT, all test scenario's pairwise comparisons - Approach III

Pairwise	Statistics of the real data				NPI-B-RP					
	Reject?	p-value	ES	Cohen's d	t-test			WMT		
					min	mean	max	min	mean	max
A vs. B	Yes	0.0003	0.226	2.041	0.914	0.931	0.949	0.869	0.895	0.919
B vs. C	Yes	0.0000	0.366	3.213	0.994	0.998	1.000	0.992	0.997	1.000
C vs. D	Yes	0.0007	0.178	1.753	0.838	0.865	0.898	0.815	0.844	0.869
D vs. E	Yes	0.0191	0.097	1.038	0.548	0.588	0.620	0.581	0.617	0.672
E vs. F	No	0.5977	-0.013	-0.115	0.886	0.909	0.932	0.914	0.934	0.953

Table B.6: Statistical and reproducibility analysis for all test scenario's pairwise comparisons - Infinite range

B.2 Reproducibility for the final decision

B.2.1 Original data

Similarly to Algorithm 1, Algorithm 2 can be applied with NPI-B with both finite and infinite intervals. This section explores how the choice of range affects reproducibility of the final decision. For finite NPI-B, the left (L) and the right (R) bounds of the support for each dose are determined similarly as for the calculation of reproducibility for separate pairwise comparisons. Approach II, introduced in Section 2.3.3, with $c = 1$ and $c = 0.5$, is presented in Figures B.1 and B.2, respectively, and the infinite range approach is presented in Figure B.3. Again, the same process has been repeated 5 times and only the first trial is displayed. Nevertheless, the outputs were similar each time and the same pattern of outcomes was shown throughout. As these illustrations show, the wider the range, the smaller the reproducibility of the final decision. Widening the range had the same and even larger effect on the results of Algorithm 2 as it did on the results of Algorithm 1. The explanation here is the same as before: a wider range creates a larger overlap between doses. The effect of this overlap is greater when more pairwise comparisons are carried out.

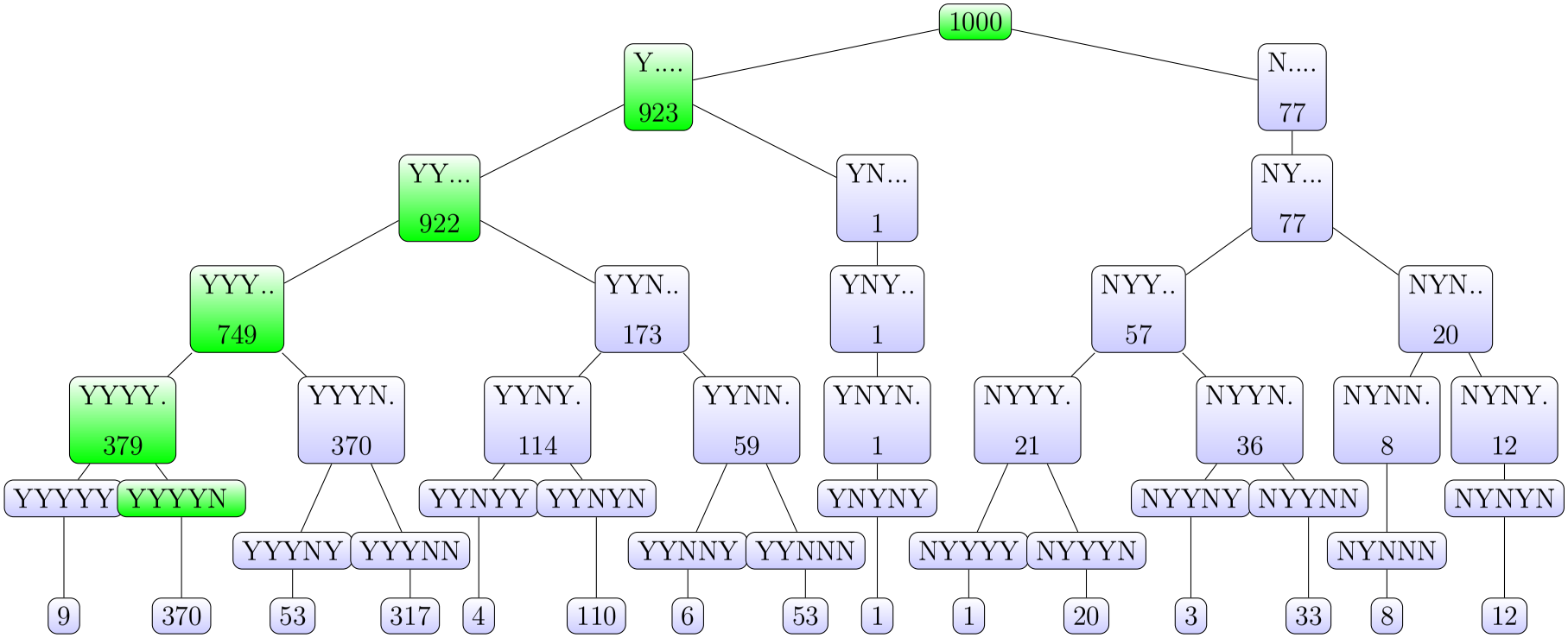


Figure B.1: Tree diagram for reproducibility of the final decision for original test scenario (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, finite range - Approach II, $c = 1$

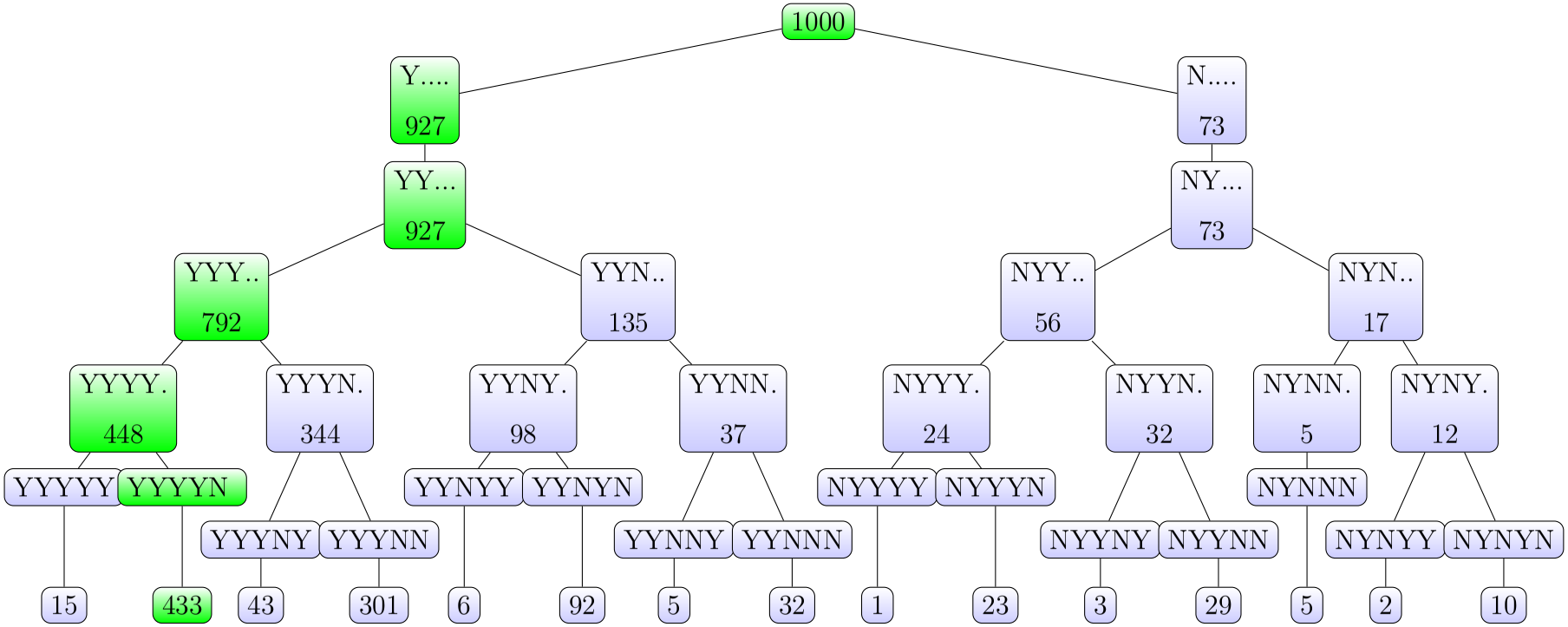


Figure B.2: Tree diagram for reproducibility of the final decision for original test scenario (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, finite range - Approach II., $c = 0.5$

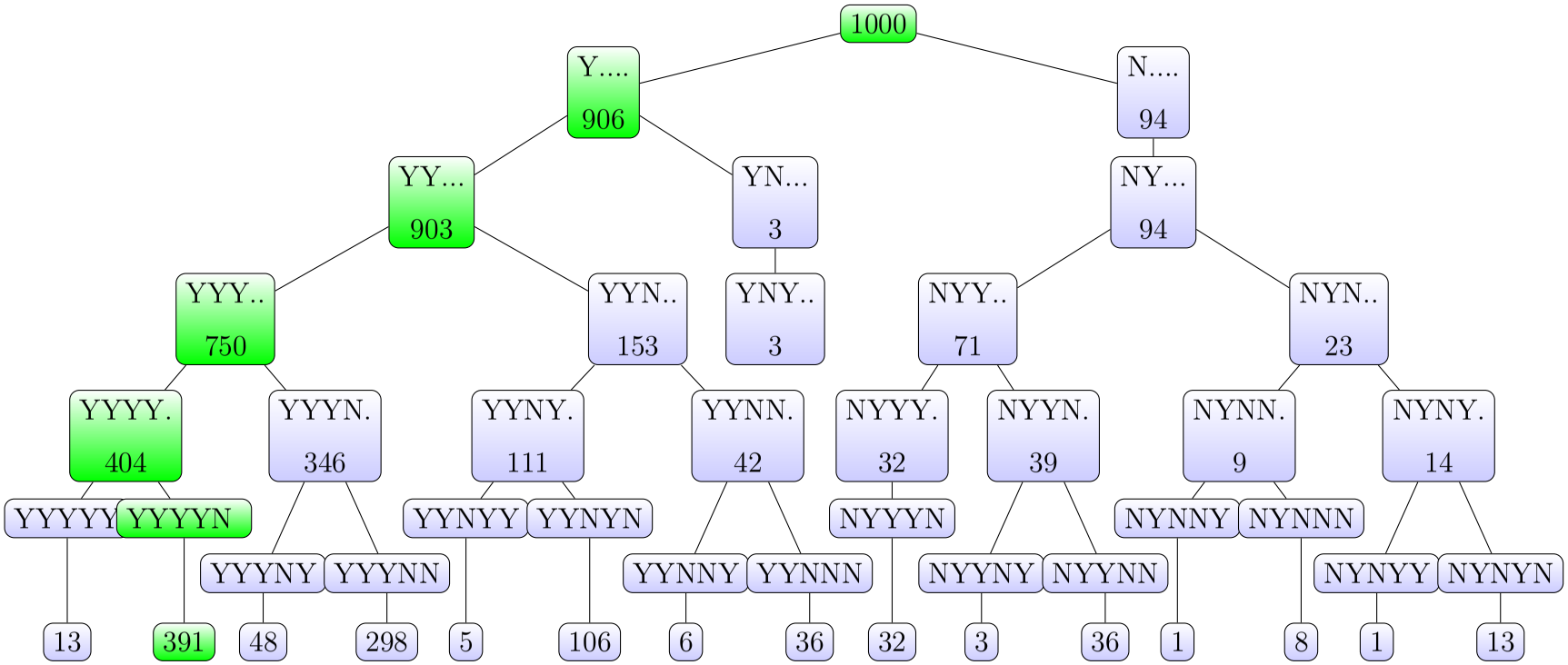


Figure B.3: Tree diagram for reproducibility of the final decision for original test scenario (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, infinite range

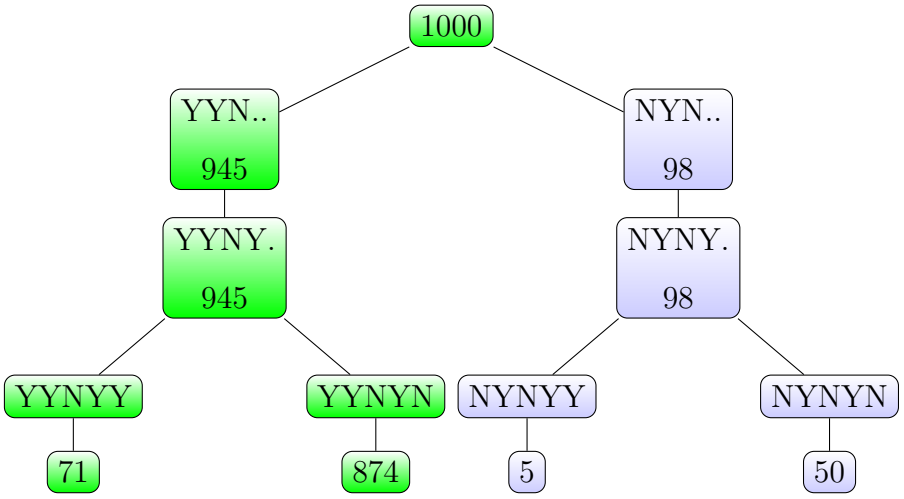


Figure B.4: Illustration of the final decision rule: Tree diagram for reproducibility of the final decision for the modified data (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, finite range - Approach II, $c = 0.5$

B.2.2 Modified data

Reproducibility of the final decision of the modified dataset, studied in Section 4.4.2, has been explored for various selections of tails. Reproducibility trees for the Approach II (for $c = 0.5, 1, 2$) are displayed in Figures B.4, B.5 and B.6, respectively, and reproducibility trees for the infinite approach in Figure B.7. Similarly to reproducibility of the final decision of the original data, widening the range leads to a lower reproducibility of the final decision for the modified data.

B.3 Outline of work leading to null findings

This section will briefly describe work carried out as part of this PhD project, which lead to null findings. The importance of presenting null and negative findings has been already highlighted in Section 3.4.2. In preclinical studies, data is not always well-behaved and Normally distributed and the right statistical analysis is not always applied to the data. This PhD work carried out a simulation study to investigate whether the reproducibility measure can detect that a wrong statistical analysis was used. An example of a wrong statistical analysis is the incorrect use of the t-test, i.e. when the t-test is applied to data

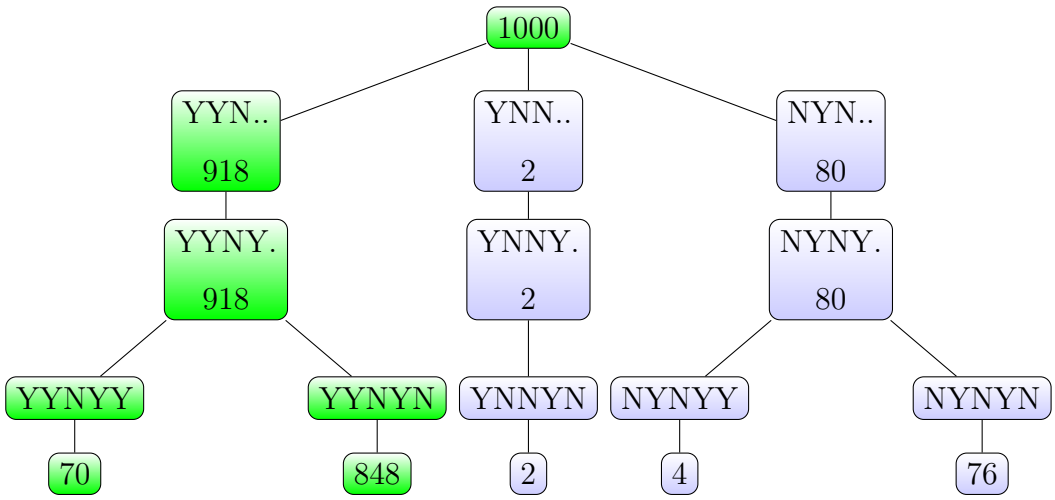


Figure B.5: Illustration of the final decision rule: Tree diagram for reproducibility of the final decision for the modified data (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, finite range - Approach II, $c = 1$

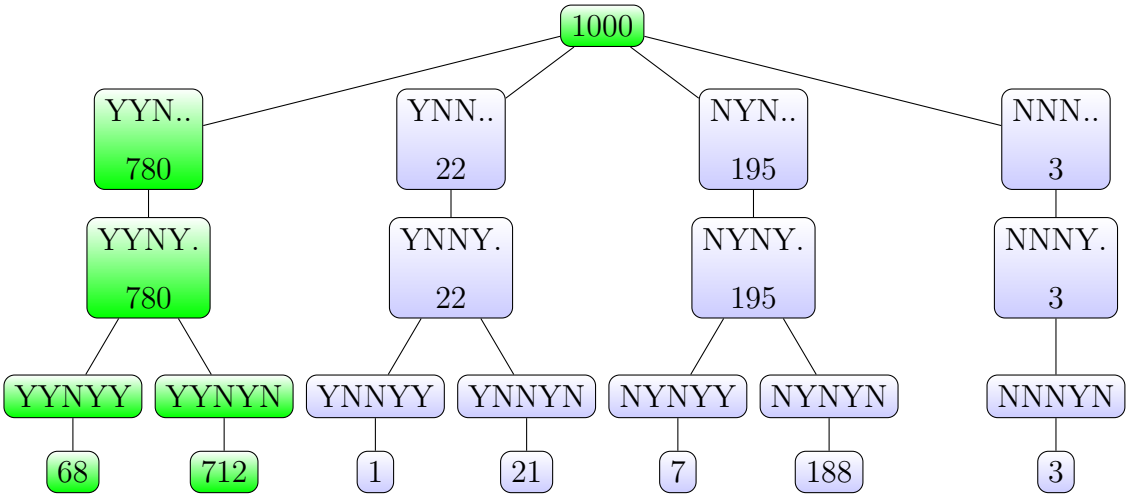


Figure B.6: Illustration of the final decision rule: Tree diagram for reproducibility of the final decision for the modified data (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, finite range - Approach II, $c = 2$

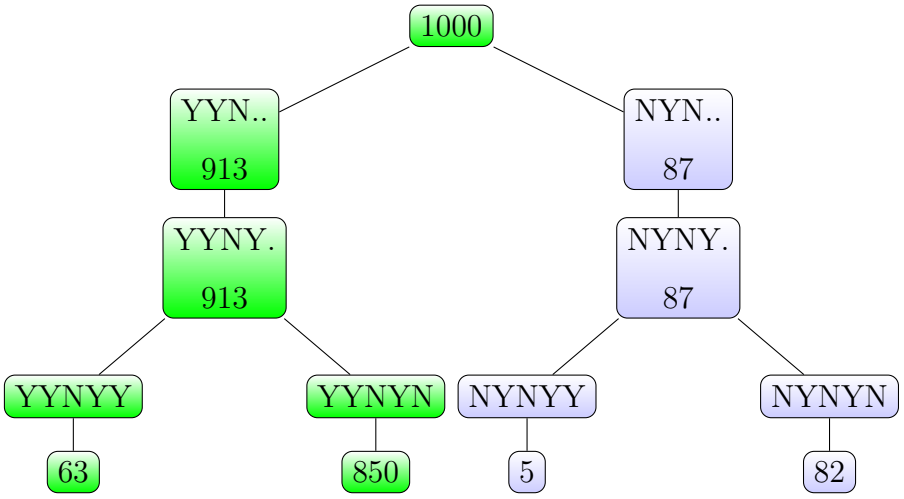


Figure B.7: Illustration of the final decision rule: Tree diagram for reproducibility of the final decision for the modified data (Outputs of Step 5 of Algorithm 6), t-test, adjusted p-value, infinite range

that do not follow Normal distribution. NPI reproducibility was not aimed at detecting the use of wrong statistical analysis, nevertheless, it has been briefly studied, as this question was of interest to practitioners.

NPI-B-RP for the t-test were calculated via Algorithm 5 for three different distributions: Normal, Lognormal and Mixed-Normal. For each distribution, for several runs, datasets were generated both under H_0 and H_1. The following details were recorded: sample size n, whether data were generated under H_0 or H_1, parameters of the underlying distribution/s, whether hypothesis was rejected or not-rejected for the t-test, Cohen's d, p-value for the t-test and NPI-B-RP for the t-test. In particular, the simulation study focused on the relationship between p-values and NPI-B-RPs for t-test and whether this is influenced by the underlying distribution of the data. The following conclusion was drawn: NPI-B-RP did not detect the use of incorrect statistical test.

Appendix C

Selected R code

C.1 R code relating to Chapter 2

R code for all bootstrap methods and an example of the R code for the comparison of bootstrap methods in their performance in both the estimation and prediction is provided here.

Efron bootstrap

```
bootstrap_efron <- function(x,m,B) {
  bootstrap_finiteinterval <- matrix(nrow = B, ncol = m)
  x <- sort(x)
  ll <- length(x)
  for (i in 1:B){ # This cycle creates B bootstraps, each containing m values
    for (j in 1:m){ # This cycle creates m new values from the original intervals
      jj<-sample(1:ll,1,prob=rep(1/ll,ll))
      new_value <- x[jj]
      bootstrap_finiteinterval[i,j] <- new_value
    }
  }
  return(bootstrap_finiteinterval)
}
```

Banks bootstrap

```
### Approach I
function_banks <- function(x, m, B) {
  x <- sort(x)
  n <- length(x)
```

```
  ## First, calculate Left (so) and Right bound of support (sn)
  distance <- vector()
  for (k in (1:(n-1))) {
    d1 <- x[(k+1)] - x[k]
    distance <- c(distance,d1)
  }
  max_distance <- max(distance)
  so <- min(x) - max_distance #  Defines x_0 for x
  sn <- max(x) + max_distance # Defines x_{n+1} for x
  x <- append(x, c(so, sn), after =  length(x)) # Add starting and ending point
  x <- sort(x)
  int_1 <- length(x)-1
  bootstrap_finiteinterval <- matrix(nrow = B, ncol = m)
  for (i in 1:B){ # This cycle creates B bootstrap samples, each containing m values
    for (j in 1:m){ # This cycle creates m new values from the original intervals
      jj <-sample(1:int_1,1,prob=rep(1/int_1,int_1))  # Sample an interval
      new_value <- runif(1, min = x[jj], max = x[jj+1]) # Sample a value in that interval
      bootstrap_finiteinterval[i,j] <- new_value
    }
  }
  return(bootstrap_finiteinterval)
}

### Approach IV
halfinfinite_banks <- function(x, m, B) {
  x <- sort(x)
  bootstrap_halfinfinite <- matrix(nrow = B, ncol = m)
  n <-length(x) # Count how many values
  int_1 <- n+1 # Count how many intervals
  for (i in 1:B){
    for (j in 1:m){
      jj<-sample(1:int_1,1,prob=rep(1/int_1,int_1)) # Sample an interval
      if (jj == 1) {
        new_value <- runif(1, min = 0, max = x[jj]) # Generate a value from the interval
    (jj-1, jj)
        bootstrap_halfinfinite[i,j] <- new_value
      }
      else if (jj == int_1){
        repeat {
          y0 <-rexp(1, rate=log(n+1)/x[n])
          if (y0 > x[n]) break
        }
        new_value <- y0
        bootstrap_halfinfinite[i,j] <- new_value
      }
      else {new_value <- runif(1, min = x[jj-1], max = x[jj]) # Generate a value from the
     interval (jj-1, jj)
```

```
      bootstrap_halfinfinite[i,j] <- new_value
      }
    }
  }
  return(bootstrap_halfinfinite)
}

### Approach V
install.packages("msm")
library(msm)
infinite_banks <- function(x, m, B) {
  x <- sort(x)
  bootstrap_infinite <- matrix(nrow = B, ncol = m)
  n <-length(x) # Count how many values
  int_1 <- n+1 # Count how many intervals
  for (i in 1:B){
    for (j in 1:m){
      jj <-sample(1:int_1,1,prob=rep(1/int_1,int_1)) # Sample an interval
      if (jj == 1) {
        mmean <- (x[1] + x[n])/2
        vvariance <- (x[n]-mmean)/qnorm(n/(int_1))
        new_value <- rtnorm(1, mean = mmean, sd = vvariance, lower = -Inf, upper = min(x)
    ) # Generate a value from the tail for  (-inf, x1)
        bootstrap_infinite[i,j] <- new_value
      } else if (jj == int_1){
        mmean <- (x[1] + x[n])/2
        vvariance <- (x[n]-mmean)/qnorm(n/(int_1))
        new_value <- rtnorm(1,mean = mmean, sd = vvariance, lower=max(x), upper=Inf) #
    Generate a value from the tail for (xn, inf)
        bootstrap_infinite[i,j] <- new_value
      }  else {new_value <- runif(1, min = x[jj-1], max = x[jj]) # generate a value from
    the interval (jj-1, jj)
      bootstrap_infinite[i,j] <- new_value
      }
    }
  }
  return(bootstrap_infinite)
}
```

NPI bootstrap

```
### Approach I
NPI_finiteI <- function(x,m,B) {
  x <-sort(x)
  n <- length(x)
  ## First, calculate Left (so) and Right bound of support (sn)
  distance <- vector()
```

```
  for (k in (1:(n-1))) {
    d1 <- x[(k+1)] - x[k]
    distance <- c(distance,d1)
  }
  max_distance <- max(distance)
  so <- min(x) - max_distance #  Defines x_0 for x
  sn <- max(x) + max_distance # Defines x_{n+1} for x
  xx <- sort(c(so,x,sn)) # Add starting and ending point
  n <- length(xx)
  lb <- matrix(c(xx[1:(n-1)],rep(NA,m)),B,n-1+m,byrow=TRUE) # Interval lower bound
  w <- matrix(c(xx[2:n]-xx[1:(n-1)],rep(NA,m)),B,n-1+m,byrow=TRUE) # Interval width
  ii <- matrix(1:B,B,2)
  for (j in 1:m) {# This cycle at one go generates step by step all B bootstrap values
    ii[,2] <- sample(n-2+j,B,replace=TRUE) # Sample an interval B times (i.e. the start
    of the interval)
    z <- runif(B) # Sample uniformly B values from 0 to 1
    lb[,n-1+j] <- lb[ii]+z*w[ii] # Calculate the value: the start of the interval + the
    width of the interval*z \in (0,1)
    w[,n-1+j] <- (1-z)*w[ii] # New interval added in
    w[ii] <- z*w[ii] # New width added in
  }
  return(lb[,n:ncol(lb)])
}

### Approach IV
halfinfinite_npi <- function(x,m,B) {
  bootstrap_halfinfinite <- matrix(nrow = B, ncol = m)
  x <- sort(x)
  for (i in 1:B){ # This cycle creates B bootstraps, each containing m values
    data <- sort(x)
    for (j in 1:m){
      n <- length(data)
      int_1 <- n+1 # Count how many intervals
      jj<-sample(1:int_1,1,prob=rep(1/int_1,int_1)) # Sample an interval
      if (jj == 1) {
        new_value <- runif(1, min = 0, max = data[jj]) # Generate a value from the
    interval (jj-1, jj)
        data <- append(data, new_value, after = jj-1)
        bootstrap_halfinfinite[i,j] <- new_value
      } else if (jj == int_1){
        repeat {
          y0 <-rexp(1, rate=log(n+1)/data[n])
          if (y0 > data[n]) break
        }
        new_value <- y0
        data <- append(data, new_value, after = n)
        bootstrap_halfinfinite[i,j] <- new_value
```

```
      } else {new_value <- runif(1, min = data[jj-1], max = data[jj]) # Generate a value
    from the interval (jj-1, jj)
        data <- append(data, new_value, after = jj-1)
        bootstrap_halfinfinite[i,j] <- new_value
      }
    }
  }
  return(bootstrap_halfinfinite)
}

### Approach V
library(msm)
infinite_npi <- function(x,m,B) {
  bootstrap_infinite <- matrix(nrow = B, ncol = m)
  x <- sort(x)
  for (i in 1:B){ # This cycle creates B bootstrap samples, each containing m values
    data <- sort(x)
    for (j in 1:m){
      n <- length(data)
      int_1 <- n+1 # Count how many intervals
      jj <-sample(1:int_1,1,prob=rep(1/int_1,int_1)) # Sample an interval
      if (jj == 1) {
        mmean <- (data[1] + data[n])/2
        vvariance <- (data[n]-mmean)/qnorm(n/(int_1))
        new_value <- rtnorm(1, mean = mmean, sd = vvariance, lower = -Inf, upper = min(
    data)) # generate a value from the tail for  (-inf, x1)
        data <- append(data, new_value, after = 0)
        bootstrap_infinite[i,j] <- new_value
      } else if (jj == int_1){
        mmean <- (data[1] + data[n])/2
        vvariance <- (data[n]-mmean)/qnorm(n/(int_1))
        new_value <- rtnorm(1,mean = mmean, sd = vvariance, lower=max(data), upper=Inf) #
     generate a value from the tail for (xn, inf)
        data <- append(data, new_value, after = n)
        bootstrap_infinite[i,j] <- new_value
      } else {new_value <- runif(1, min = data[jj-1], max = data[jj]) # Generate a value
    from the interval (jj-1, jj)
        data <- append(data, new_value, after = jj-1)
        bootstrap_infinite[i,j] <- new_value
      }
    }
  }
  return(bootstrap_infinite)
}
```

Hutson bootstrap

```
### ($-\infty , \infty$)
hutson_bootstrap <- function(x, m, B) {
  x <- sort(x)
  bootstrap_infinite <- matrix(nrow = B, ncol = m)
  n <-length(x) # Count how many values
  int_1 <- n+1 # Count how many intervals
  for (i in 1:B){
    for (j in 1:m){
      jj <- runif(1)
      if (jj <= (1/int_1)) {
        new_value <- x[1]+(x[2]-x[1])*log(int_1*jj)
        bootstrap_infinite[i,j] <- new_value
      }
      else if (jj >= (n/int_1)){
        new_value <- x[n]-(x[n]-x[(n-1)])*log(int_1*(1-jj))
        bootstrap_infinite[i,j] <- new_value
      }
      else{
        xx <- int_1*jj
        floor_xx <- floor(xx)
        eps <- xx - floor_xx
        new_value <- (1-eps)*x[floor_xx]+eps*x[(floor_xx+1)]
        bootstrap_infinite[i,j] <- new_value
      }
    }
  }
  return(bootstrap_infinite)
}

### [$0, \infty$)
hutson_bootstrap_from0 <- function(x, m, B) {
  x <- sort(x)
  bootstrap_infinite <- matrix(nrow = B, ncol = m)
  n <-length(x) # Count how many values
  int_1 <- n+1 # Count how many intervals
  for (i in 1:B){
    for (j in 1:m){
      jj <- runif(1)
      if (jj <= (1/int_1)) {
        xx <- int_1*jj
        floor_xx <- floor(xx)
        eps <- xx - floor_xx
        new_value <- x[1]*eps
        bootstrap_infinite[i,j] <- new_value
      }
```

```
      else if (jj >= (n/int_1)){
        new_value <- x[n]-(x[n]-x[(n-1)])*log(int_1*(1-jj))
        bootstrap_infinite[i,j] <- new_value
      }
      else{
        xx <- int_1*jj
        floor_xx <- floor(xx)
        eps <- xx - floor_xx
        new_value <- (1-eps)*x[floor_xx]+eps*x[(floor_xx+1)]
        bootstrap_infinite[i,j] <- new_value
      }
    }
  }
  return(bootstrap_infinite)
}
```

Kernel bootstrap

```
### Bandwidth for kernel density estimate
h.window <-function(x, n, window=1){
  if (window==1) {
    wind.weight<-(1.06)*sd(x)*n^(-1/5)
  } else if (window==2) {
    wind.weight<-(.79)*IQR(x)*n^(-1/5)
  } else if (window==3) {
    wind.weight <- min(sd(x),(.90)*(IQR(x)/1.34)*n^(-1/5))
  }
  return(wind.weight)
}

kernel_b<-function(x, n, B, window=1){
  # x is the original original dataset, n is the bootstrap sample size,
  # B is the number of bootstrap samples generated, window is the selected bandwidth
  x.length <-n*B
  x.unif <-matrix(sample(x, size=x.length, replace=T),
                  nrow=B)
  x.h <-h.window(x,n,window)
  epsilon <-matrix(rnorm(x.length), nrow=B)
  y <-x.unif+x.h*epsilon
  return(y)}
```

Illustration of Algorithm 1

The following R code runs Algorithm 1 for Normally distributed data, percentile confidence intervals are applied; finite Approach I is applied to Banks-B and NPI-B.

```
chi_square_coverage_normal_finiteI_manyseeds <- function(seed=20,m=1,s=1,sample_sizes=c
    (4,6,8,10),statist="median",B=1000,N=1000){
  ### true statistic
  if (statist=="q1") {
    true_q1 <- m-0.67448*s
  } else if (statist=="q3") {
    true_q3 <- m+0.67448*s
  } else if (statist=="variance") {
    true_variance <- s^2
  } else if (statist=="iqr") {
    true_iqr <- 1.34896*s
  }
  function_single_boot <- function(boot = "NPI",m,s,seeds,S1=S1,n,statist="median",B,N) {
    ## Now we need 1000 simulated data,
    confidence_regions <- data.frame(CR10=logical(),
                                     CR20=logical(),
                                     CR30=logical(),
                                     CR40=logical(),
                                     CR50=logical(),
                                     CR60=logical(),
                                     CR70=logical(),
                                     CR80=logical(),
                                     CR90=logical(),
                                     stringsAsFactors=FALSE)
    # for simulated data, create 1000 bootstraps, and calculate 1000 $\hat{\theta}$
    for (i in 1:N) {
      ss <- S1[,i]
      if (boot=="NPI") {
        set.seed(seeds)
        SS <- NPI_finiteI(ss,n,B)
      } else if (boot=="Efron") {
        set.seed(seeds)
        SS <- bootstrap_efron(ss,n,B)
      } else if (boot=="Banks"){
        set.seed(seeds)
        SS <- function_banks(ss,n,B)
      } else if (boot=="Hutson"){
        set.seed(seeds)
        SS <- hutson_bootstrap(ss,n,B)
      } else {print("ERROR: no such bootstrap programmed")
      }
      ## Compute statistics of the bootstrapped samples
```

```
 if (statist=="median") {
  boot_stat <- apply(SS, 1, median)
} else if (statist=="q1") {
  quantile1 <- function(x){
    return(quantile(x, c(0.25)))
  }
  boot_stat <- apply(SS, 1, quantile1)
} else if (statist=="q3") {
  quantile3 <- function(x){
    return(quantile(x, c(0.75)))
  }
  boot_stat <- apply(SS, 1, quantile3)
} else if (statist=="mean") {
  boot_stat <- apply(SS, 1, mean)
} else if (statist=="variance") {
  boot_stat <- apply(SS, 1, var)
} else if (statist=="iqr") {
  boot_stat <- apply(SS, 1, IQR)
} else {
  print("ERROR: no such statistic programmed")
}
boot_stat <- sort(boot_stat)
## Compute (1-2alpha)\% intervals (q_alpha,q_(1-alpha))
# for true $\theta$
# at different convidence levels (10\%, 20\%, ..., 100\%)
# 10\% - alpha=0.45, CI=(0.45,0.55)
CI10 <- c(boot_stat[(B*0.45)],boot_stat[(B*0.55)])
# 20\% - alpha=0.40, CI=(0.40,0.60)
CI20 <- c(boot_stat[(B*0.4)],boot_stat[(B*0.6)])
# 30\% - alpha=0.35, CI=(0.35,0.65)
CI30 <- c(boot_stat[(B*0.35)],boot_stat[(B*0.65)])
# 40\% - alpha=0.30, CI=(0.30,0.70)
CI40 <- c(boot_stat[(B*0.3)],boot_stat[(B*0.7)])
# 50\% - alpha=0.25, CI=(0.25,0.75)
CI50 <- c(boot_stat[(B*0.25)],boot_stat[(B*0.75)])
# 60\% - alpha=0.20, CI=(0.20,0.80)
CI60 <- c(boot_stat[(B*0.2)],boot_stat[(B*0.8)])
# 70\% - alpha=0.15, CI=(0.15,0.85)
CI70 <- c(boot_stat[(B*0.15)],boot_stat[(B*0.85)])
# 80\% - alpha=0.10, CI=(0.10,0.90)
CI80 <- c(boot_stat[(B*0.1)],boot_stat[(B*0.9)])
# 90\% - alpha=0.05, CI=(0.05,0.95)
CI90 <- c(boot_stat[(B*0.05)],boot_stat[(B*0.95)])
#CI95 <- c(boot_stat[25],boot_stat[975])
### Which CI include the true $\theta$?
if (statist=="median") {
```

```
    new <- c(CI10[1] <= m & CI10[2] >= m, CI20[1] <= m & CI20[2] >= m, CI30[1] <= m &
 CI30[2] >= m,CI40[1] <= m & CI40[2] >= m,CI50[1] <= m & CI50[2] >= m,CI60[1] <= m &
CI60[2] >= m,CI70[1] <= m & CI70[2] >= m,CI80[1] <= m & CI80[2] >= m,CI90[1] <= m &
CI90[2] >= m)
  } else if (statist=="q1") {
    new <- c(CI10[1] <= true_q1 & CI10[2] >= true_q1, CI20[1] <= true_q1 & CI20[2] >=
 true_q1, CI30[1] <= true_q1 & CI30[2] >= true_q1,CI40[1] <= true_q1 & CI40[2] >=
true_q1,CI50[1] <= true_q1 & CI50[2] >= true_q1,CI60[1] <= true_q1 & CI60[2] >= true_
q1,CI70[1] <= true_q1 & CI70[2] >= true_q1,CI80[1] <= true_q1 & CI80[2] >= true_q1,
CI90[1] <= true_q1 & CI90[2] >= true_q1)
  } else if (statist=="q3") {
    new <- c(CI10[1] <= true_q3 & CI10[2] >= true_q3, CI20[1] <= true_q3 & CI20[2] >=
 true_q3, CI30[1] <= true_q3 & CI30[2] >= true_q3,CI40[1] <= true_q3 & CI40[2] >=
true_q3,CI50[1] <= true_q3 & CI50[2] >= true_q3,CI60[1] <= true_q3 & CI60[2] >= true_
q3,CI70[1] <= true_q3 & CI70[2] >= true_q3,CI80[1] <= true_q3 & CI80[2] >= true_q3,
CI90[1] <= true_q3 & CI90[2] >= true_q3)
  } else if (statist=="mean") {
    new <- c(CI10[1] <= (m) & CI10[2] >= (m), CI20[1] <= (m) & CI20[2] >= (m), CI30
[1] <= (m) & CI30[2] >= (m),CI40[1] <= (m) & CI40[2] >= (m),CI50[1] <= (m) & CI50[2]
>= (m),CI60[1] <= (m) & CI60[2] >= (m),CI70[1] <= (m) & CI70[2] >= (m),CI80[1] <= (m)
 & CI80[2] >= (m),CI90[1] <= (m) & CI90[2] >= (m))
  } else if (statist=="variance") {
    new <- c(CI10[1] <= true_variance & CI10[2] >= true_variance, CI20[1] <= true_
variance & CI20[2] >= true_variance, CI30[1] <= true_variance & CI30[2] >= true_
variance,CI40[1] <= true_variance & CI40[2] >= true_variance,CI50[1] <= true_variance
 & CI50[2] >= true_variance,CI60[1] <= true_variance & CI60[2] >= true_variance,CI70
[1] <= true_variance & CI70[2] >= true_variance,CI80[1] <= true_variance & CI80[2] >=
 true_variance,CI90[1] <= true_variance & CI90[2] >= true_variance)
  } else if (statist=="iqr") {
    new <- c(CI10[1] <= true_iqr & CI10[2] >= true_iqr, CI20[1] <= true_iqr & CI20[2]
 >= true_iqr, CI30[1] <= true_iqr & CI30[2] >= true_iqr,CI40[1] <= true_iqr & CI40[2]
 >= true_iqr,CI50[1] <= true_iqr & CI50[2] >= true_iqr,CI60[1] <= true_iqr & CI60[2]
>= true_iqr,CI70[1] <= true_iqr & CI70[2] >= true_iqr,CI80[1] <= true_iqr & CI80[2]
>= true_iqr,CI90[1] <= true_iqr & CI90[2] >= true_iqr)
  } else {
    print("Error: no such statistic defined")
  }
  confidence_regions <-rbind(confidence_regions,do.call(data.frame,setNames(as.list(
new), names(confidence_regions))))
}
## What is the proportion of CI with true $\theta$?
## This gives the true actual probabilities for the true $\theta$
prop_sumCR10 <- sum(confidence_regions$CR10)/N
prop_sumCR20 <- sum(confidence_regions$CR20)/N
prop_sumCR30 <-sum(confidence_regions$CR30)/N
prop_sumCR40 <-sum(confidence_regions$CR40)/N
prop_sumCR50 <-sum(confidence_regions$CR50)/N
```

```
  prop_sumCR60 <-sum(confidence_regions$CR60)/N
  prop_sumCR70 <-sum(confidence_regions$CR70)/N
  prop_sumCR80 <-sum(confidence_regions$CR80)/N
  prop_sumCR90 <-sum(confidence_regions$CR90)/N
  ## This gives the actual coverage for the true  $\theta$
  coverageCR10 <-prop_sumCR10
  coverageCR20 <-prop_sumCR20 - prop_sumCR10
  coverageCR30 <-prop_sumCR30 - prop_sumCR20
  coverageCR40 <-prop_sumCR40 - prop_sumCR30
  coverageCR50 <-prop_sumCR50 - prop_sumCR40
  coverageCR60 <-prop_sumCR60 - prop_sumCR50
  coverageCR70 <-prop_sumCR70 - prop_sumCR60
  coverageCR80 <-prop_sumCR80 - prop_sumCR70
  coverageCR90 <-prop_sumCR90 - prop_sumCR80
  cr1_9 <- sum(c(coverageCR10,coverageCR20,coverageCR30,coverageCR40,coverageCR50,
  coverageCR60,coverageCR70,coverageCR80,coverageCR90))
  coverageCR100 <- 1 - cr1_9
  coverage <- N*c(coverageCR10,coverageCR20,coverageCR30,coverageCR40,coverageCR50,
  coverageCR60,coverageCR70,coverageCR80,coverageCR90,coverageCR100)
  ## Apply the chi-square test of goodness of fit
  chi_square <- as.numeric(chisq.test(coverage, p = rep(1/10,10))$statistic)
  return(as.vector(c(n,seeds,boot,(prop_sumCR90*100),chi_square)))
}
table <- data.frame(sample_size=integer(),
                    seed=integer(),
                    boot=factor(),
                    cov90p=numeric(),
                    chi2=numeric(),
                    stringsAsFactors=FALSE)
 for (k in 1:length(sample_sizes)) {
 n <- sample_sizes[k]
 set.seed(1)
 S1 <- replicate(N,rnorm(n,m,s))
      for (j in 1:seed) {
              npi_b <- function_single_boot(boot = "NPI",m,s,seeds=j,S1=S1,n,statist=
  statist,B=B,N=N)
              table <-rbind(table,do.call(data.frame,setNames(as.list(npi_b), names(
  table))))
               banks_b <- function_single_boot(boot = "Banks",m,s,seeds=j,S1=S1,n,
  statist=statist,B=B,N=N)
              table <-rbind(table,do.call(data.frame,setNames(as.list(banks_b), names(
  table))))
              efron_b <- function_single_boot(boot = "Efron",m,s,seeds=j,S1=S1,n,
  statist=statist,B=B,N=N)
              table <-rbind(table,do.call(data.frame,setNames(as.list(efron_b), names(
  table))))
```

```
                hutson_b <- function_single_boot(boot = "Hutson",m,s,seeds=j,S1=S1,n,
    statist=statist,B=B,N=N)
                table <-rbind(table,do.call(data.frame,setNames(as.list(hutson_b), names(
    table))))
        }
    }
  return(table)
}
```

The following R code runs Algorithm 1 for Lognormally distributed data, BC_a confidence intervals are applied; half-finite Approach V is applied to Banks-B and NPI-B.

```
chi_square_coverage_lognormal_halfinfinite_manyseeds_bca <- function(seed=20,m=1,s=1,
    sample_sizes=c(4,6,8,10),statist="median",B=1000,N=1000){
  # Lognormal parameters:
  meanlog=log(m^2/sqrt(m^2+s^2))
  sdlog=sqrt(log(s^2/m^2+1))
  ### True statistics:
  if (statist=="q1") {
    true_q1 <- exp(meanlog+sdlog*qnorm(0.25))
  } else if (statist=="q3") {
    true_q3 <- exp(meanlog+sdlog*qnorm(0.75))
  } else if (statist=="variance") {
    true_variance <- s^2
  } else if (statist=="iqr") {
    true_iqr <- exp(meanlog+sdlog*qnorm(0.75))-exp(meanlog+sdlog*qnorm(0.25))
  } else if (statist=="median") {
    true_median <- exp(meanlog+sdlog*qnorm(0.50))
  }
  function_single_boot <- function(boot = "NPI",m,s,seeds,S1=S1,n,statist="median",B,N) {
    ## Now we need 1000 simulated data,
    confidence_regions <- data.frame(CR10=logical(),
                                     CR20=logical(),
                                     CR30=logical(),
                                     CR40=logical(),
                                     CR50=logical(),
                                     CR60=logical(),
                                     CR70=logical(),
                                     CR80=logical(),
                                     CR90=logical(),
                                     stringsAsFactors=FALSE)
    # For simulated data, create 1000 bootstraps, and from those calculate 1000 $\hat{\
    theta}$
      for (i in 1:N) {
      ss <- S1[,i]
      if (boot=="NPI") {
        set.seed(seeds)
        SS <- halfinfinite_npi(ss,n,B=1000)
```

```
  } else if (boot=="Efron") {
    set.seed(seeds)
    SS <- bootstrap_efron(ss,n,1000)
  } else if (boot=="Banks"){
    set.seed(seeds)
    SS <- halfinfinite_banks(ss,n,1000)
  } else if (boot=="Hutson"){
    set.seed(seeds)
    SS <- hutson_bootstrap_from0(ss,n,1000)
  } else {
    print("ERROR: no such bootstrap programmed")
  }
  ## Compute statistics of the bootstrapped samples
  if (statist=="median") {
    boot_stat <- apply(SS, 1, median)
    T_n <-median(ss) # original estimate of T_n
  } else if (statist=="q1") {
    quantile1 <- function(x){
      return(quantile(x, c(0.25)))
    }
    boot_stat <- apply(SS, 1, quantile1)
    T_n <-quantile1(ss) # original estimate of T_n
  } else if (statist=="q3") {
    quantile3 <- function(x){
      return(quantile(x, c(0.75)))
    }
    boot_stat <- apply(SS, 1, quantile3)
    T_n <-quantile3(ss) # original estimate of T_n
  } else if (statist=="mean") {
    boot_stat <- apply(SS, 1, mean)
    T_n <-mean(ss) # original estimate of T_n
  } else if (statist=="variance") {
    boot_stat <- apply(SS, 1, var)
    T_n <-var(ss) # original estimate of T_n
  } else if (statist=="iqr") {
    boot_stat <- apply(SS, 1, IQR)
    T_n <-IQR(ss) # original estimate of T_n
  }
  boot_stat <- sort(boot_stat)
  # Now calculate bias correciton z_0:
  z_0 <- qnorm((sum(boot_stat<T_n)/B))
  # Find jacknife values of a statistic T_n. Let x_i =(x_1,x_2,...,x_(i-1),x_(i+1)
,...x_n)
  # be the jacknife smaple which is the original sample with the ith observation x_i
deleted.
  jacknife <- vector()
  for (k in 1:(length(ss))) {
```

```
    jacki <- ss[-k]
    if (statist=="median") {
      T_n_i <-median(jacki) # original estimate of T_n_i
    } else if (statist=="q1") {
      T_n_i <-quantile1(jacki) # original estimate of T_n_i
    } else if (statist=="q3") {
      T_n_i <-quantile3(jacki) # original estimate of T_n_i
    } else if (statist=="mean") {
      T_n_i <-mean(jacki) # original estimate of T_n_i
    } else if (statist=="variance") {
      T_n_i <-var(jacki) # original estimate of T_n_i
    } else if (statist=="iqr") {
      T_n_i <-IQR(jacki) # original estimate of T_n_i
    } else {
      print("ERROR: no such statistic programmed")
    }
    jacknife <- c(jacknife,T_n_i)
  }
  T_n. <- sum(jacknife)/length(ss)
  ## To calculate acceleration alpha
  a_hat <- sum((T_n.-jacknife)^3)/(6*(sum((T_n.-jacknife)^2))^(3/2))
  ## Compute (1-2alpha)\% intervals (q_alpha,q_(1-alpha))
  # for true statistic
  # at different confidence levels (10\%, 20\%, ..., 100\%)
  # 10\% - alpha=0.45, CI=(0.45,0.55)
  alpha_1_10 <- pnorm((z_0+(z_0+qnorm(0.45))/(1-a_hat*(z_0+qnorm(0.45)))))
  if (alpha_1_10<0.001) {alpha_1_10=0.001}
  alpha_2_10 <- pnorm((z_0+(z_0+qnorm(0.55))/(1-a_hat*(z_0+qnorm(0.55)))))
  CI10 <- c(boot_stat[alpha_1_10 *1000],boot_stat[alpha_2_10*1000])
  # 20\% - alpha=0.40, CI=(0.40,0.60)
  alpha_1_20 <- pnorm((z_0+(z_0+qnorm(0.40))/(1-a_hat*(z_0+qnorm(0.40)))))
  if (alpha_1_20<0.001) {alpha_1_20=0.001}
  alpha_2_20 <- pnorm((z_0+(z_0+qnorm(0.60))/(1-a_hat*(z_0+qnorm(0.60)))))
  CI20 <- c(boot_stat[alpha_1_20 *1000],boot_stat[alpha_2_20*1000])
  # 30\% - alpha=0.35, CI=(0.35,0.65)
  alpha_1_30 <- pnorm((z_0+(z_0+qnorm(0.35))/(1-a_hat*(z_0+qnorm(0.35)))))
  if (alpha_1_30<0.001) {alpha_1_30=0.001}
  alpha_2_30 <- pnorm((z_0+(z_0+qnorm(0.65))/(1-a_hat*(z_0+qnorm(0.65)))))
  CI30 <- c(boot_stat[alpha_1_30 *1000],boot_stat[alpha_2_30*1000])
  # 40\% - alpha=0.30, CI=(0.30,0.70)
  alpha_1_40 <- pnorm((z_0+(z_0+qnorm(0.30))/(1-a_hat*(z_0+qnorm(0.30)))))
  if (alpha_1_40<0.001) {alpha_1_40=0.001}
  alpha_2_40 <- pnorm((z_0+(z_0+qnorm(0.70))/(1-a_hat*(z_0+qnorm(0.70)))))
  CI40 <- c(boot_stat[alpha_1_40 *1000],boot_stat[alpha_2_40*1000])
  # 50\% - alpha=0.25, CI=(0.25,0.75)
  alpha_1_50 <- pnorm((z_0+(z_0+qnorm(0.25))/(1-a_hat*(z_0+qnorm(0.25)))))
  if (alpha_1_50<0.001) {alpha_1_50=0.001}
```

```
    alpha_2_50 <- pnorm((z_0+(z_0+qnorm(0.75))/(1-a_hat*(z_0+qnorm(0.75)))))
    CI50 <- c(boot_stat[alpha_1_50 *1000],boot_stat[alpha_2_50*1000])
    # 60\% - alpha=0.20, CI=(0.20,0.80)
    alpha_1_60 <- pnorm((z_0+(z_0+qnorm(0.20))/(1-a_hat*(z_0+qnorm(0.20)))))
    if (alpha_1_60<0.001) {alpha_1_60=0.001}
    alpha_2_60 <- pnorm((z_0+(z_0+qnorm(0.80))/(1-a_hat*(z_0+qnorm(0.80)))))
    CI60 <- c(boot_stat[alpha_1_60 *1000],boot_stat[alpha_2_60*1000])
    # 70\% - alpha=0.15, CI=(0.15,0.85)
    alpha_1_70 <- pnorm((z_0+(z_0+qnorm(0.15))/(1-a_hat*(z_0+qnorm(0.15)))))
    if (alpha_1_70<0.001) {alpha_1_70=0.001}
    alpha_2_70 <- pnorm((z_0+(z_0+qnorm(0.85))/(1-a_hat*(z_0+qnorm(0.85)))))
    CI70 <- c(boot_stat[alpha_1_70 *1000],boot_stat[alpha_2_70*1000])
    # 80\% - alpha=0.10, CI=(0.10,0.90)
    alpha_1_80 <- pnorm((z_0+(z_0+qnorm(0.10))/(1-a_hat*(z_0+qnorm(0.10)))))
    if (alpha_1_80<0.001) {alpha_1_80=0.001}
    alpha_2_80 <- pnorm((z_0+(z_0+qnorm(0.90))/(1-a_hat*(z_0+qnorm(0.90)))))
    CI80 <- c(boot_stat[alpha_1_80 *1000],boot_stat[alpha_2_80*1000])
    # 90\% - alpha=0.05, CI=(0.05,0.95)
    alpha_1_90 <- pnorm((z_0+(z_0+qnorm(0.05))/(1-a_hat*(z_0+qnorm(0.05)))))
    if (alpha_1_90<0.001) {alpha_1_90=0.001}
    alpha_2_90 <- pnorm((z_0+(z_0+qnorm(0.95))/(1-a_hat*(z_0+qnorm(0.95)))))
    CI90 <- c(boot_stat[alpha_1_90 *1000],boot_stat[alpha_2_90*1000])
    ### Which CI include the true $\hat{\theta}$?
    if (statist=="median") {
      new <- c(CI10[1] <= true_median & CI10[2] >= true_median , CI20[1] <= true_median
    & CI20[2] >= true_median, CI30[1] <= true_median  & CI30[2] >= true_median,CI40[1]
<= true_median  & CI40[2] >= true_median,CI50[1] <= true_median & CI50[2] >= true_
median,CI60[1] <= true_median  & CI60[2] >= true_median,CI70[1] <= true_median  &
CI70[2] >= true_median,CI80[1] <= true_median  & CI80[2] >= true_median ,CI90[1] <=
true_median  & CI90[2] >= true_median )
    } else if (statist=="q1") {
      new <- c(CI10[1] <= true_q1 & CI10[2] >= true_q1, CI20[1] <= true_q1 & CI20[2] >=
  true_q1, CI30[1] <= true_q1 & CI30[2] >= true_q1,CI40[1] <= true_q1 & CI40[2] >=
true_q1,CI50[1] <= true_q1 & CI50[2] >= true_q1,CI60[1] <= true_q1 & CI60[2] >= true_
q1,CI70[1] <= true_q1 & CI70[2] >= true_q1,CI80[1] <= true_q1 & CI80[2] >= true_q1,
CI90[1] <= true_q1 & CI90[2] >= true_q1)
    } else if (statist=="q3") {
      new <- c(CI10[1] <= true_q3 & CI10[2] >= true_q3, CI20[1] <= true_q3 & CI20[2] >=
  true_q3, CI30[1] <= true_q3 & CI30[2] >= true_q3,CI40[1] <= true_q3 & CI40[2] >=
true_q3,CI50[1] <= true_q3 & CI50[2] >= true_q3,CI60[1] <= true_q3 & CI60[2] >= true_
q3,CI70[1] <= true_q3 & CI70[2] >= true_q3,CI80[1] <= true_q3 & CI80[2] >= true_q3,
CI90[1] <= true_q3 & CI90[2] >= true_q3)
    } else if (statist=="mean") {
      new <- c(CI10[1] <= (m) & CI10[2] >= (m), CI20[1] <= (m) & CI20[2] >= (m), CI30
[1] <= (m) & CI30[2] >= (m),CI40[1] <= (m) & CI40[2] >= (m),CI50[1] <= (m) & CI50[2]
>= (m),CI60[1] <= (m) & CI60[2] >= (m),CI70[1] <= (m) & CI70[2] >= (m),CI80[1] <= (m)
  & CI80[2] >= (m),CI90[1] <= (m) & CI90[2] >= (m))
```

```
  } else if (statist=="variance") {
    new <- c(CI10[1] <= true_variance & CI10[2] >= true_variance, CI20[1] <= true_
variance & CI20[2] >= true_variance, CI30[1] <= true_variance & CI30[2] >= true_
variance,CI40[1] <= true_variance & CI40[2] >= true_variance,CI50[1] <= true_variance
 & CI50[2] >= true_variance,CI60[1] <= true_variance & CI60[2] >= true_variance,CI70
[1] <= true_variance & CI70[2] >= true_variance,CI80[1] <= true_variance & CI80[2] >=
 true_variance,CI90[1] <= true_variance & CI90[2] >= true_variance)
  } else if (statist=="iqr") {
    new <- c(CI10[1] <= true_iqr & CI10[2] >= true_iqr, CI20[1] <= true_iqr & CI20[2]
 >= true_iqr, CI30[1] <= true_iqr & CI30[2] >= true_iqr,CI40[1] <= true_iqr & CI40[2]
 >= true_iqr,CI50[1] <= true_iqr & CI50[2] >= true_iqr,CI60[1] <= true_iqr & CI60[2]
>= true_iqr,CI70[1] <= true_iqr & CI70[2] >= true_iqr,CI80[1] <= true_iqr & CI80[2]
>= true_iqr,CI90[1] <= true_iqr & CI90[2] >= true_iqr)
  } else {
    print("Error: no such statistic defined")
  }
  confidence_regions <-rbind(confidence_regions,do.call(data.frame,setNames(as.list(
new), names(confidence_regions))))
}
## What is the proportion of CI with true $\theta$?
## This gives the true actual probabilities for the true  $\theta$
prop_sumCR10 <- sum(confidence_regions$CR10)/N
prop_sumCR20 <- sum(confidence_regions$CR20)/N
prop_sumCR30 <-sum(confidence_regions$CR30)/N
prop_sumCR40 <-sum(confidence_regions$CR40)/N
prop_sumCR50 <-sum(confidence_regions$CR50)/N
prop_sumCR60 <-sum(confidence_regions$CR60)/N
prop_sumCR70 <-sum(confidence_regions$CR70)/N
prop_sumCR80 <-sum(confidence_regions$CR80)/N
prop_sumCR90 <-sum(confidence_regions$CR90)/N
## This gives the actual coverage for the true  $\theta$
coverageCR10 <-prop_sumCR10
coverageCR20 <-prop_sumCR20 - prop_sumCR10
coverageCR30 <-prop_sumCR30 - prop_sumCR20
coverageCR40 <-prop_sumCR40 - prop_sumCR30
coverageCR50 <-prop_sumCR50 - prop_sumCR40
coverageCR60 <-prop_sumCR60 - prop_sumCR50
coverageCR70 <-prop_sumCR70 - prop_sumCR60
coverageCR80 <-prop_sumCR80 - prop_sumCR70
coverageCR90 <-prop_sumCR90 - prop_sumCR80
cr1_9 <- sum(c(coverageCR10,coverageCR20,coverageCR30,coverageCR40,coverageCR50,
coverageCR60,coverageCR70,coverageCR80,coverageCR90))
coverageCR100 <- 1 - cr1_9
coverage <- N*c(coverageCR10,coverageCR20,coverageCR30,coverageCR40,coverageCR50,
coverageCR60,coverageCR70,coverageCR80,coverageCR90,coverageCR100)
## Apply the chi-square test of goodness of fit
chi_square <- as.numeric(chisq.test(coverage, p = rep(1/10,10))$statistic)
```

```
    return(as.vector(c(n,seeds,boot,(prop_sumCR90*100),chi_square)))
  }
  table <- data.frame(sample_size=numeric(),
                      seed=integer(),
                      boot=factor(),
                      cov90p=numeric(),
                      chi2=numeric(),
                      stringsAsFactors=FALSE)
  for (k in 1:length(sample_sizes)) {
    n <- sample_sizes[k]
    set.seed(1)
    S1 <- replicate(N,rlnorm(n,meanlog=meanlog,sdlog=sdlog))
    for (j in 1:seed) {
      npi_b <- function_single_boot(boot = "NPI",m,s,seeds=j,S1=S1,n,statist=statist,B=B,
    N=N)
      table <-rbind(table,do.call(data.frame,setNames(as.list(npi_b), names(table))))
      banks_b <- function_single_boot(boot = "Banks",m,s,seeds=j,S1=S1,n,statist=statist,
    B=B,N=N)
      table <-rbind(table,do.call(data.frame,setNames(as.list(banks_b), names(table))))
      efron_b <- function_single_boot(boot = "Efron",m,s,seeds=j,S1=S1,n,statist=statist,
    B=B,N=N)
      table <-rbind(table,do.call(data.frame,setNames(as.list(efron_b), names(table))))
      hutson_b <- function_single_boot(boot = "Hutson",m,s,seeds=j,S1=S1,n,statist=
    statist,B=B,N=N)
      table <-rbind(table,do.call(data.frame,setNames(as.list(hutson_b), names(table))))
    }
  }
  return(table)
}
```

Illustration of Algorithm 2

```
quantile3 <- function(x){
  return(quantile(x, c(0.75)))
}

quantile1 <- function(x){
  return(quantile(x, c(0.25)))
}

prediction_performance_normal_finite <- function(m=1,s=1,sample_size_options=c
    (4,6,8,10,20),N=1000,B=1000,seeds=20,alp=0.05,statistic="mean") {
    ### Table where we record outputs:
    table <- data.frame(sample_size=integer(),
                        run=integer(),
                        boot=factor(),
```

```
                        sum_lies=integer(),
                        stringsAsFactors=FALSE)
bootstraps <-c("NPI","Banks","Efron","Hutson")
len<-length(sample_size_options)
for (k in 1:seeds){ # For each run
  for (i in 1:len) { # For each sample size
    n <- sample_size_options[i]
    set.seed(k)
    # For each run, we draw $2N$ samples and let $X_1, X_2, ... Y_N$
    # be the actual samples and $X_{N+1}, X_{N+2}, ... Y_{2N}$ be the future samples.
    S <- replicate((2*N),rnorm(n,m,s))
    # current samples
    Sx <- S[,1:N]
    # future samples
    Sy <- S[,(N+1):(2*N)]
    if (statistic=="mean") {
      stat_y <- apply(Sy,MARGIN=2,FUN=mean)
    } else if (statistic=="median") {
      stat_y <- apply(Sy,MARGIN=2,FUN=median)
    } else if (statistic=="variance") {
      stat_y <- apply(Sy,MARGIN=2,FUN=var)
    } else if (statistic=="q1") {
      stat_y <- apply(Sy,MARGIN=2,FUN=quantile1)
    } else if (statistic=="q3") {
      stat_y <- apply(Sy,MARGIN=2,FUN=quantile3)
    } else if (statistic=="iqr") {
      stat_y <- apply(Sy,MARGIN=2,FUN=IQR)
    }
    for (j in 1:4) {
      boot=bootstraps[j]
      does_lie <- vector()
      for (l in 1:N) {
        ss = Sx[,l]
        if (boot=="NPI") {
          set.seed(k)
          SS <- NPI_finiteI(ss,n,B)
        } else if (boot=="Efron") {
          set.seed(k)
          SS <- bootstrap_efron(ss,n,B)
        } else if (boot=="Banks"){
          set.seed(k)
          SS <- function_banks(ss,n,B)
        } else if (boot=="Hutson"){
          set.seed(k)
          SS <- hutson_bootstrap(ss,n,B)
        } else {
          print("ERROR: no such bootstrap programmed")
```

```
            }
            if (statistic=="mean") {
              stat_boot <- sort(apply(SS,MARGIN=1,FUN=mean))
            } else if (statistic=="variance") {
              stat_boot <- sort(apply(SS,MARGIN=1,FUN=var))
            } else if (statistic=="q1") {
              stat_boot <- sort(apply(SS, 1, quantile1))
            } else if  (statistic=="q3") {
              stat_boot <- sort(apply(SS, 1, quantile3))
            } else if (statistic=="median") {
              stat_boot <- sort(apply(SS,MARGIN=1,FUN=median))
            } else if (statistic=="iqr") {
              stat_boot <- sort(apply(SS,MARGIN=1,FUN=IQR))
            }
            lower_PI <- stat_boot[(alp*B)]
            upper_PI <- stat_boot[((1-alp)*B)]
            lies <- lower_PI <= stat_y[i] & upper_PI >= stat_y[i]
            does_lie <- append(does_lie,lies)
          }
          sum_lies <- sum(does_lie)
          new <- c(n,k,boot,sum_lies)
          table <-rbind(table,do.call(data.frame,setNames(as.list(new), names(table))))
        }
      }
    }
    return(table)
}
```

C.2 R code relating to Chapter 4

R code for calculation of NPI-B-RP for the t-test and the growth rate inhibition significance test, and of estimates of NPI-RP for the WMT and the t-test is provided in this section.

NPI-B-RP for the t-test (Algorithm 5)

```
#### NPI-B-RP for the $t$-test, finite Approach I
# NPI-B-RP is calculated for the Students one-sided t-test (equal variance t-test) which
    compares whether the mean for dose 1 is bigger than for dose 2
## The below code can be adjusted for the WMT and for different ranges
reproducibility_t_test_finiteI <- function(dose1,dose2) {
  dose1 <- sort(dose1) # Sort data in increasing order for dose 1
  n <- length(dose1) # Calculate sample size of dose 1
  # Record all distances between adjacent points for dose 1
```

```
max <- vector()
for (k in (1:(n-1))) {
  max1 <- dose1[(k+1)] - dose1[k]
  max <- c(max,max1)
}
max_max <- max(max)  # Record maximal distance between adjacent points for dose 1
dose2<- sort(dose2) # Sort data in increasing order for dose 2
n2 <- length(dose2) # Calculate sample size of dose 2
# Record all distances between adjacent points for dose 2
max2 <- vector()
for (k in (1:(n2-1))) {
  max_2 <- dose2[(k+1)] - dose2[k]
  max2 <- c(max2,max_2)
}
max_max2 <- max(max2) # Record maximal distance between adjacent points for dose 2
so1 <- min(dose1) - max_max # Defines x_0 for dose 1
sn1 <- max(dose1) + max_max # Defines x_{n+1} for dose 1
so2 <- min(dose2) - max_max2 # Defines x_0 for dose 2
sn2 <- max(dose2) + max_max2 # Defines x_{n+1} for dose 2
one_times_step_NPI_finiteI <- function(dose1,dose2,so1,sn1,so2,sn2) {
  function_comparison_NPI_finiteI <- function(dose1,dose2,so1,sn1,so2,sn2) {
    # This function calculates p-value between 2 bootstrapped samples
    NPI_finite <- function(x, so, sn) { # This function creates one bootstrap sample
      m <- length(x)
      x <- append(x, c(so, sn), after =  length(x)) # Add starting and ending point
      x <- sort(x) # Sort data in increasing order for the sample
      boot <- vector()
      for (j in 1:m){ # This cycle creates m new values from the original intervals
        int_1 <- length(x) - 1
        jj <-sample(1:int_1,1,prob=rep(1/int_1,int_1))  # Sample an interval
        new_value <- runif(1, min = x[jj], max = x[jj+1]) # Sample a value in that
  interval
        x <- append(x, new_value, after = jj) # Add this sampled value to the set of
  values
        boot <- c(boot, new_value)
      }
      return(t(boot))
    }
    x1 <- NPI_finite(dose1,so1,sn1) # Create bootstrap sample for dose 1
    y1 <- NPI_finite(dose2,so2,sn2) # Create bootstrap sample for dose 2
    return(t.test(x1,y1, alternative = "greater", paired = FALSE, var.equal = TRUE)$p.
  value)
    # Calculate p-value between 2 bootstrapped samples
  }
  total <- sum(replicate(1000, function_comparison_NPI_finiteI(dose1,dose2,so1,sn1,so2,
  sn2)) <=0.05) # Repeat 1000 (N) times Step 2 of the Algorithm
  # Calculate how many times we got the same decision as was the original decision:
```

```
    if (t.test(dose1,dose2, alternative = "greater", paired = FALSE, var.equal = TRUE)$p.
    value < 0.05) {
      rp <- total/1000}
    else {rp <- 1 - total/1000}
    return(rp)
  }
  # The below line performs Steps 2-4 of the algorithm h (100) times
  output <-replicate(100, one_times_step_NPI_finiteI(dose1, dose2, so1,sn1,so2,sn2))
  return(c(min(output),mean(output),max(output)))
}
```

Reproducibility of the final decision (Algorithm 6)

```
function_final_decision_all_combinations_t_test_finiteI <- function(dose1,dose2,dose3,
    dose4,dose5,dose6) {
  n <- length(dose1) # Calculate sample size of dose 1
  dose1 <- sort(dose1) # Sort data in increasing order for dose 1
  # Record all distances between adjacent points for dose 1
  avl <- vector()
  for (k in (1:(n-1))) {
    avl1 <- dose1[(k+1)] - dose1[k]
    avl <- c(avl,avl1)
  }
  max1 <- max(avl) # Record maximal distance between adjacent points for dose 1
  n2 <- length(dose2) # Calculate sample size of dose 2
  dose2 <- sort(dose2) # Sort data in increasing order for dose 2
  # Record all distances between adjacent points for dose 2
  avl2 <- vector()
  for (k in (1:(n2-1))) {
    avl_2 <- dose2[(k+1)] - dose2[k]
    avl2 <- c(avl2,avl_2)
  }
  max2 <- max(avl2) # Record maximal distance between adjacent points for dose 2
  n3 <- length(dose3) # Calculate sample size of dose 3
  dose3 <- sort(dose3) # Sort data in increasing order for dose 3
  # Record all distances between adjacent points for dose 3
  avl3 <- vector()
  for (k in (1:(n3-1))) {
    avl_3 <- dose3[(k+1)] - dose3[k]
    avl3 <- c(avl3,avl_3)
  }
  max3 <- max(avl3) # Record maximal distance between adjacent points for dose 3
  n4 <- length(dose4) # Calculate sample size of dose 4
  dose4 <- sort(dose4) # Sort data in increasing order for dose 4
  # Record all distances between adjacent points for dose 4
  avl4 <- vector()
  for (k in (1:(n4-1))) {
```

```
  avl_4 <- dose4[(k+1)] - dose4[k]
  avl4 <- c(avl4,avl_4)
}
max4 <- max(avl4) # Record maximal distance between adjacent points for dose 4
n5 <- length(dose5) # Calculate sample size of dose 5
dose5 <- sort(dose5) # Sort data in increasing order for dose 5
# Record all distances between adjacent points for dose 5
avl5 <- vector()
for (k in (1:(n5-1))) {
  avl_5 <- dose5[(k+1)] - dose5[k]
  avl5 <- c(avl5,avl_5)
}
max5 <- max(avl5) # Record maximal distance between adjacent points for dose 5
n6 <- length(dose6) # Calculate sample size of dose 6
dose6 <- sort(dose6) # Sort data in increasing order for dose 6
# Record all distances between adjacent points for dose
avl6 <- vector()
for (k in (1:(n6-1))) {
  avl_6 <- dose6[(k+1)] - dose6[k]
  avl6 <- c(avl6,avl_6)
}
max6 <- max(avl6) # Record maximal distance between adjacent points for dose 6
so1 = min(dose1)-max1 # Defines x_0 for dose 1
sn1 = max(dose1)+max1 # Defines x_{n+1} for dose 1
so2 = min(dose2)-max2 # Defines x_0 for dose 2
sn2 = max(dose2)+max2 # Defines x_{n+1} for dose 2
so3 = min(dose3)-max3 # Defines x_0 for dose 3
sn3 = max(dose3)+max3 # Defines x_{n+1} for dose 3
so4 = min(dose4)-max4 # Defines x_0 for dose 4
sn4 = max(dose4)+max4 # Defines x_{n+1} for dose 4
so5 = min(dose5)-max5 # Defines x_0 for dose 5
sn5 = max(dose5)+max5 # Defines x_{n+1} for dose 5
so6 = min(dose6)-max6 # Defines x_0 for dose 6
sn6 = max(dose6)+max6 # Defines x_{n+1} for dose 6
first <- matrix(NA, nrow=1,ncol=10)
for (i in 1:10) { # Note: 10 can be changed into a different number, depending on how
  many outputs we wants
  NPI_finiteI <- function(x,s0=-Inf,sn=Inf,m,B) { # This function creates B bootstrap
  sample
    xx <- sort(c(s0,x,sn)) # Now it contains min and max point
    n <- length(xx)
    lb <- matrix(c(xx[1:(n-1)],rep(NA,m)),B,n-1+m,byrow=TRUE) # Interval lower bound
    w <- matrix(c(xx[2:n]-xx[1:(n-1)],rep(NA,m)),B,n-1+m,byrow=TRUE) # Interval width
    ii <- matrix(1:B,B,2)
    for (j in 1:m) {# This cycle at one go generates step by step all B bootstrap
  values
      ii[,2] <- sample(n-2+j,B,replace=TRUE) # Sample an interval B times (i.e. the
```

```
start of the interval)
    z <- runif(B) # Sample uniformly B values from 0 to 1
    lb[,n-1+j] <- lb[ii]+z*w[ii] # Calculate the value: the start of the interval +
the width of the interval*z \in (0,1)
    w[,n-1+j] <- (1-z)*w[ii] # New interval added in
    w[ii] <- z*w[ii] # New width added in
  }
  return(lb[,n:ncol(lb)])
}
# Calculate how many new points should be created for each dose
m1 <- length(dose1) # Calculate sample size of dose 1
m2 <- length(dose2) # Calculate sample size of dose 2
m3 <- length(dose3) # Calculate sample size of dose 3
m4 <- length(dose4) # Calculate sample size of dose 4
m5 <- length(dose5) # Calculate sample size of dose 5
m6 <- length(dose6) # Calculate sample size of dose 6
# Create new set of data points for each dose N (1000) times
new_d1 <- NPI_finiteI(dose1,so1,sn1,m1,1000)
new_d2 <- NPI_finiteI(dose2,so2,sn2,m2,1000)
new_d3 <- NPI_finiteI(dose3,so3,sn3,m3,1000)
new_d4 <- NPI_finiteI(dose4,so4,sn4,m4,1000)
new_d5 <- NPI_finiteI(dose5,so5,sn5,m5,1000)
new_d6 <- NPI_finiteI(dose6,so6,sn6,m6,1000)
# Create an empty matrix in which you will record the findigs each time
total_conclusion <- matrix(NA,1,5)
for (i in 1:1000) {
  # each time do the pairwise comparisons for the bootstrapped samples
  p1 <- t.test(new_d1[i,], new_d2[i,],alternative = "greater", paired = FALSE, var.
equal = TRUE)$p.value
  p2 <- t.test(new_d2[i,],new_d3[i,], alternative = "greater", paired = FALSE, var.
equal = TRUE)$p.value
  p3 <- t.test(new_d3[i,],new_d4[i,], alternative = "greater", paired = FALSE, var.
equal = TRUE)$p.value
  p4 <- t.test(new_d4[i,],new_d5[i,], alternative = "greater", paired = FALSE, var.
equal = TRUE)$p.value
  p5 <- t.test(new_d5[i,],new_d6[i,], alternative = "greater", paired = FALSE, var.
equal = TRUE)$p.value
  # collect the p values in a vector p.raw
  p.raw_ttest <- c(p1, p2, p3, p4, p5)
  # Since we run 5 tests simultaneously, we adjust the p-values for multiple testing
  # using the Benjamini & Hochberg (1995) procedure
  # and see whether it is below 0.05 or not
  conclusion <- p.adjust(p.raw_ttest, method = "BH", n = length(p.raw_ttest)) < 0.05
  conclusion <- matrix(conclusion,1,5) # collect in a vector (in a matrix form)
  total_conclusion <-rbind(total_conclusion,conclusion) # Each time, add a line to
our matrix, recording results from each attempt
}
```

```
    total_conclusion <- total_conclusion[-1,] # remove the first line (which is not
    needed)
    # Create a frequency table of all the possible combinations of test outcomes recorded
     in Step 3 of the algorithm
    total <- apply(total_conclusion, 1, function(x) paste(x, collapse=".")  )
    kk <- sort(table(total))
    print(kk)
  }
}
```

Sampling of orderings for the WMT (Algorithm 7)

```
### For upper tail one-sided WMT, i.e. testing whether data for dose1 is shifted to the
    left of data for dose2
### The below function is used for test scenario from Section 4.2
npi_rp_function_srs_independent_max <- function(dose1,dose2,ntt,za,reject = T) { # Here
    the test is whether dose1 is smaller than dose2
  # ntt stands for number of orderings chosen
  # za is the rank sum test statistic Z, read from Tables (see Hollander and Wolfe)
  # This value is specific for particular n and m
  # Values available only for m>=0 and n>=10
  # 1) Find Left and Right bound of support for each dose - using finite I
  m <- length(dose1)
  distance <- vector()
  dose1<-sort(dose1)
  for (k in (1:(m-1))) {
    d1 <- dose1[(k+1)] - dose1[k]
    distance <- c(distance,d1)
  }
  max_value <- max(distance) # Max distance for dose1
  n <- length(dose2)
  dose2 <-sort(dose2)
  distance2 <- vector()
  for (k in (1:(n-1))) {
    d2 <- dose2[(k+1)] - dose2[k]
    distance2 <- c(distance2,d2)
  }
  max_value2 <- max(distance2) # Max distance for dose2
  L <- dose1[1]-max_value
  R <- dose1[m]+max_value
  dose1 <- sort(c(L,dose1,R))
  L2 <- dose2[1]-max_value2
  R2 <- dose2[n]+max_value2
  dose2 <- sort(c(L2,dose2,R2))
  ## 2) Get n* possible orderings of the future observations for both doses
  srs <- function(n) { # n stands for number of original points, ntt stands for number of
     orderings sampled
```

```
  x.index <- sample(1:(2*n), n) # Draw n values from 1:(2*n) (withouth replacement)
  x.index <- sort(x.index)
  x.index1<- c(0,  x.index, 2*n+1) # We add to this two more value: 0 and m+n+1 (0 and
  the number of intervals at the end)
  ss<-diff(x.index1)-1  # to get the diff between the orderings
  print(ss)
}
XX <- replicate(ntt,srs((length(dose1)-2)))
YY <- replicate(ntt,srs((length(dose2)-2)))
# From the proof page 95 (Bin thesis), put the future at x_(j) or y_(j-1) for the lower
   and the other way around for the upper
XL<- as.data.frame(apply(XX, 2, FUN= function(X) rep(dose1[-length(dose1)],X))) # - the
   last value (gets rid of it)
XU<- as.data.frame(apply(XX, 2, FUN= function(X) rep(dose1[-1],X))) # - the first value
   (gets rid of it)
YL<- as.data.frame(apply(YY, 2, FUN= function(Y) rep(dose2[-length(dose2)],Y)))
YU<- as.data.frame(apply(YY, 2, FUN= function(Y) rep(dose2[-1],Y)) )
# calculate minimum value of the rank sum
lrp_sapply <- vector()
for (i in 1:ntt) {
  a <- wilcox.test(as.vector(t(YL[i])),as.vector(t(XU[i])),exact=F)$statistic + (n*(n
  +1)/2) # This calculated W statistic for dose2 (for lower repr.)
  lrp_sapply <- c(lrp_sapply,a)
}
# Calculate maximum value of the rank sum
urp_sapply <- vector()
for (j in 1:ntt) {
  b <- wilcox.test(as.vector(t(YU[j])),as.vector(t(XL[j])),exact=F)$statistic + (n*(n
  +1)/2) # This gives out the rank sum statistic (i.e. Sum of ranks of Y). # This
  calculated W statistic for dose2 (for lower repr.)
  urp_sapply <- c(urp_sapply,b)
}
if (reject == T) {
  lower <- sum(lrp_sapply>=za)/length(lrp_sapply) # Lower RP for H_0 rejection
  upper <- sum(urp_sapply>=za)/length(urp_sapply) # Upper RP for H_0 rejection
} else {
  lower = 1 - sum(urp_sapply>=za)/length(urp_sapply) # Lower RP for H_0 non-rejection
  upper = 1 - sum(lrp_sapply>=za)/length(lrp_sapply) # Upper RP for H_0 non-rejection
}
output_rp <- c(lower,upper)
## Calculate confidence intervals for both the lower and upper RP estimate
confidence_intervals <- function(output_rp,ntt) {
  confidence <- vector()
  for (i in 1:2) {
    x <- output_rp[i]
    x_l <- x-1.960*sqrt((x*(1-x))/ntt)
    x_u <- x+1.960*sqrt((x*(1-x))/ntt)
```

```
      confidence <- c(confidence,x_l,x_u)
    }
    return(confidence)
  }
  confidences <-  confidence_intervals(output_rp,ntt)
  output <- c(output_rp,confidences)
  print(output)
}## The output of the function is: lower RP estimate, bigger RP estimate, CI for lower RP
     estimate, CI for upper RP estimate

#### The function needs to be adjusted for large sample sizes (over m=10 x n=10)
#### Used for test scenarios in Section 4.7
npi_rp_function_srs_independent_max_large <- function(dose1,dose2,ntt,za=1.645,reject = T
    ) { # Here the test is whether dose1 is smaller than dose2 # za_0.05=1.645
  # ntt stands for number of orderings chosen
  # Dose 1 is the Y variable and Dose 2 is the X variable
  # 1) Find Left and Right bound of support for each dose - using finite max approach
  m <- length(dose1)
  distance <- vector()
  dose1<-sort(dose1)
  for (k in (1:(m-1))) {
    d1 <- dose1[(k+1)] - dose1[k]
    distance <- c(distance,d1)
  }
  max_value <- max(distance) # Max distance for dose1
  n <- length(dose2)
  dose2<-sort(dose2)
  distance2 <- vector()
  for (k in (1:(n-1))) {
    d2 <- dose2[(k+1)] - dose2[k]
    distance2 <- c(distance2,d2)
  }
  max_value2 <- max(distance2) # Max distance for dose2
  L <- dose1[1]-max_value
  R <- dose1[m]+max_value
  dose1 <- sort(c(L,dose1,R))
  L2 <- dose2[1]-max_value2
  R2 <- dose2[n]+max_value2
  dose2 <- sort(c(L2,dose2,R2))
  ## 2) Get n* possible orderings of the future observations for both doses
  srs <- function(n) { # n stands for number of original points, ntt stands for number of
     orderings sampled
    x.index <- sample(1:(2*n), n) # Draw n values from 1:(2*n) (without replacement)
    x.index <- sort(x.index)
    x.index1<- c(0,  x.index, 2*n+1) # We add to this two more value: 0 and m+n+1 (0 and
    the number of intervals at the end)
    ss<-diff(x.index1)-1  # to get the diff between the orderings
```

```
  print(ss)
}
XX <- replicate(ntt,srs((length(dose1)-2)))
YY <- replicate(ntt,srs((length(dose2)-2)))
# From the proof page 95 (BinHimd's thesis), put the future at xj or yj-1 for the lower
  and the other way around for the upper
XL<- as.data.frame(apply(XX, 2, FUN= function(X) rep(dose1[-length(dose1)],X))) # - the
  last value (gets rid of it)
XU<- as.data.frame(apply(XX, 2, FUN= function(X) rep(dose1[-1],X))) # - the first value
  (gets rid of it)
YL<- as.data.frame(apply(YY, 2, FUN= function(Y) rep(dose2[-length(dose2)],Y)))
YU<- as.data.frame(apply(YY, 2, FUN= function(Y) rep(dose2[-1],Y)) )
# Calculate minimum value of the rank sum
lrp_sapply <- vector()
for (i in 1:ntt) {
  a <- wilcox.test(as.vector(t(YL[i])),as.vector(t(XU[i])),exact=FALSE)$statistic + (n*
  (n+1)/2) # Function with column from x and column from y as inputs
  lrp_sapply <- c(lrp_sapply,a)
}
# Calculate maximum value of the rank sum
urp_sapply <- vector()
for (j in 1:ntt) {
  b <- wilcox.test(as.vector(t(YU[j])),as.vector(t(XL[j])),exact=FALSE)$statistic + (n*
  (n+1)/2) # This gives out the rank sum statistic (i.e. Sum of ranks of Y). Function
  with column from x and column from y as inputs
  urp_sapply <- c(urp_sapply,b)
}
# For the Normal approximation:
E_0 <- n*(m+n+1)/2
var_0 <- m*n*(m+n+1)/12
# Look at each set of orderings and apply large sample approximation to get W_0, then
  see if this value
# is bigger than z_alpha
if (reject == T) {
  lower <- sum(((lrp_sapply-E_0)/sqrt(var_0))>=za)/length(lrp_sapply) # Lower RP for H_
  0 rejection
  upper <- sum(((urp_sapply-E_0)/sqrt(var_0))>=za)/length(urp_sapply) # Upper RP for H_
  0 rejection
} else {
  lower = 1 - sum(((urp_sapply-E_0)/sqrt(var_0))>=za)/length(urp_sapply)
  upper = 1 - sum(((lrp_sapply-E_0)/sqrt(var_0))>=za)/length(lrp_sapply)
}
output_rp <- c(lower,upper)
# calculate CI for both the lower and upper estimate of RP
confidence_intervals <- function(output_rp,ntt) {
  confidence <- vector()
  for (i in 1:2) {
```

```
      x <- output_rp[i]
      x_l <- x-1.960*sqrt((x*(1-x))/ntt)
      x_u <- x+1.960*sqrt((x*(1-x))/ntt)
      confidence <- c(confidence,x_l,x_u)
    }
    return(confidence)
  }
  confidences <-  confidence_intervals(output_rp,ntt)
  output <- c(output_rp,confidences) ## The output of the function is: lower RP estimate,
     bigger RP estimate, CI for lower RP estimate, CI for upper RP estimate
  print(output)
}
```

Sampling of orderings for the t-test (Algorithm 8)

```
#### NUMERATOR APPROACH
### Upper sided t-test - equal variance, assume \alpha=0.05
### For lower bound: we put all xs to the left and all ys to the right
### For upper bound: we put all xs to the right adn all ys to the left
sampling_t_test_numerator <- function(dose1,dose2,ntt,t,reject = T) { # Here the test is
    whether dose1 is bigger than dose2
  # ntt stands for number of orderings chosen
  # t stands for the critical t-value # This value is specific for particular n and m
  # 1) Find Left and Right bound of support for each dose - using finite I approach
  m <- length(dose1)
  dose1 <- sort(dose1)
  distance <- vector()
  for (k in (1:(m-1))) {
    dist1 <- dose1[(k+1)] - dose1[k]
    distance <- c(distance,dist1)
  }
  max_value <- max(distance) # Max distance for dose1
  dose2 <- sort(dose2)
  n <- length(dose2)
  distance2 <- vector()
  for (k in (1:(n-1))) {
    dist2 <- dose2[(k+1)] - dose2[k]
    distance2 <- c(distance2,dist2)
  }
  max_value2 <- max(distance2) # Max distance for dose2
  L <- dose1[1]-max_value
  R <- dose1[m]+max_value
  dose1 <- sort(c(L,dose1,R))
  L2 <- dose2[1]-max_value2
  R2 <- dose2[n]+max_value2
  dose2 <- sort(c(L2,dose2,R2))
  ## 2) Sample n* possible orderings of the future observations for both doses
```

```
# Function srs samples on ordering
srs <- function(n) { # n stands for number of original points, ntt stands for number of
    orderings sampled
  x.index <- sample(1:(2*n), n) # Draw n values from 1:(2*n) (without replacement)
  x.index <- sort(x.index)
  x.index1<- c(0,  x.index, 2*n+1) # We add to this two more value: 0 and m+n+1 (0 and
  the number of intervals at the end)
  ss<-diff(x.index1)-1  # to get the diff between the orderings
  print(ss)
}
# Sample n* for each dose
XX <- replicate(ntt,srs((length(dose1)-2)))
YY <- replicate(ntt,srs((length(dose2)-2)))
# put both dose1 and dose2 to both left and right
XL<- as.data.frame(apply(XX, 2, FUN= function(X) rep(dose1[-length(dose1)],X))) # - the
    last value (gets rid of it)
XU<- as.data.frame(apply(XX, 2, FUN= function(X) rep(dose1[-1],X))) # - the first value
    (gets rid of it)
YL<- as.data.frame(apply(YY, 2, FUN= function(Y) rep(dose2[-length(dose2)],Y)))
YU<- as.data.frame(apply(YY, 2, FUN= function(Y) rep(dose2[-1],Y)) )
### calculate minimum value of the rank sum for each pair of orderings
lrp_sapply <- vector()
for (i in 1:ntt) {
  a <- t.test(as.vector(t(XL[i])),as.vector(t(YU[i])), alternative = ("greater"),
  paired = FALSE, var.equal = TRUE)$statistic
  lrp_sapply <- c(lrp_sapply,a)
}
### Calculate the maximum value of the rank sum for each pair of orderings
urp_sapply <- vector()
for (j in 1:ntt) {
  b <- t.test(as.vector(t(XU[j])),as.vector(t(YL[j])), alternative = ("greater"),
  paired = FALSE, var.equal = TRUE)$statistic
  urp_sapply <- c(urp_sapply,b)
}
### Calculate the mean of t_l's and t_u's
mean_l <- mean(lrp_sapply)
mean_u <- mean(urp_sapply)
if (reject == T) {
  lower <- sum(lrp_sapply>=t)/length(lrp_sapply) # Lower RP for H0 rejection
  upper <- sum(urp_sapply>=t)/length(urp_sapply) # Upper RP for H0 rejection # same
  results as we got above
} else {
  lower = 1 - sum(urp_sapply>=t)/length(urp_sapply) # Lower RP for H0 non-rejection
  upper = 1 - sum(lrp_sapply>=t)/length(lrp_sapply) # Upper RP for H0 non-rejection
}
output_rp <- c(lower,upper)
output <- c(output_rp,mean_l,mean_u) ## The outputs of the function are: lower RP
```

```
    estimate, bigger RP estimate mean of t_l's, mean of t_u's
  print(output)
}

#### DENOMINATOR APPROACH
### For lower bound: large variance: first half to left, second half to right
### For upper bound: small variance: first half to right, second half to left
sampling_t_test_denominator <- function(dose1,dose2,ntt=1000,t,reject=T) {
  # The below function samples one ordering for a given original sample
  srs <- function(n) { # n stands for number of original points, ntt stands for number of
     orderings sampled
    x.index <- sample(1:(2*n), n) # Draw n values from 1:(2*n) (withouth replacement)
    x.index <- sort(x.index)
    x.index1<- c(0,  x.index, 2*n+1) # We add to this two more value: 0 and m+n+1 (0 and
    the number of intervals at the end)
    ss<-diff(x.index1)-1  # to get the diff between the orderings
    print(ss)
  }
  dose1 <-sort(dose1)
  m <- length(dose1)
  distance <- vector()
  for (k in (1:(m-1))) {
    dist1 <- dose1[(k+1)] - dose1[k]
    distance <- c(distance,dist1)
  }
  max_value <- max(distance) # max distance for dose1
  dose2 <-sort(dose2)
  n <- length(dose2)
  distance2 <- vector()
  for (k in (1:(n-1))) {
    dist2 <- dose2[(k+1)] - dose2[k]
    distance2 <- c(distance2,dist2)
  }
  max_value2 <- max(max2) # max distance for dose2
  L <- dose1[1]-max_value
  R <- dose1[m]+max_value
  dose1 <- sort(c(L,dose1,R))
  L2 <- dose2[1]-max_value2
  R2 <- dose2[n]+max_value2
  dose2 <- sort(c(L2,dose2,R2))
  ### This function calculates lower and upper t-value for one sampling of orderings
  one_cycle <- function(dose1,dose2,t=2.120) {
    XX <- replicate(1,srs((length(dose1)-2)))
    YY <- replicate(1,srs((length(dose2)-2)))
    XX_small<- XX[1:(length(XX)/2)]
    XX_big <- XX[-(1:(length(XX)/2))]
    XX_small_b <- c(XX_small,rep(0,times=length(XX_big)))
```

```
  XX_big_b <- c(rep(0,times=length(XX_small)),XX_big)
  YY_small<- YY[1:(length(YY)/2)]
  YY_big <- YY[-(1:(length(YY)/2))]
  YY_small_b <- c(YY_small,rep(0,times=length(YY_big)))
  YY_big_b <- c(rep(0,times=length(YY_small)),YY_big)
  XL_small <- rep(dose1[-length(dose1)],XX_small_b)
  XU_small <- rep(dose1[-1],XX_small_b)
  XL_big <- rep(dose1[-length(dose1)],XX_big_b)
  XU_big <- rep(dose1[-1],XX_big_b)
  XL <- c(XL_small, XU_big)
  XU <- c(XU_small, XL_big)
  YY_small<- YY[1:(length(YY)/2)]
  YY_big <- YY[-(1:(length(YY)/2))]
  YY_small_b <- c(YY_small,rep(0,times=length(YY_big)))
  YY_big_b <- c(rep(0,times=length(YY_small)),YY_big)
  YY_small<- YY[1:(length(YY)/2)]
  YY_big <- YY[-(1:(length(YY)/2))]
  YY_small_b <- c(YY_small,rep(0,times=length(YY_big)))
  YY_big_b <- c(rep(0,times=length(YY_small)),YY_big)
  YL_small <- rep(dose2[-length(dose2)],YY_small_b)
  YU_small <- rep(dose2[-1],YY_small_b)
  YL_big <- rep(dose2[-length(dose2)],YY_big_b)
  YU_big <- rep(dose2[-1],YY_big_b)
  YL <- c(YL_small, YU_big) # more spread -> larger variance
  YU <- c(YU_small, YL_big) # less spread -> smaller
  # Calculate minimum value of the rank sum
  a <- t.test(XL,YL, alternative = ("greater"), paired = FALSE, var.equal = TRUE)$
  statistic
  # Calculate maximum value of the rank sum
  b <- t.test(XU,YU, alternative = ("greater"), paired = FALSE, var.equal = TRUE)$
  statistic
  if (reject == T) {
    aa <- a >= t
    bb <- b >= t
  } else {
    aa <- a < t
    bb <- b < t
  }
  return(list("lower"=aa,"upper"=bb,"tl"=a,"tu"=b))
}
low <- vector()
high <- vector()
t_low <- vector()
t_high <- vector()
for (i in 1:ntt) {
  al <- one_cycle(dose1,dose2)
  low <- c(low,al$lower)
```

```
    high <-c(high,al$upper)
    t_low <- c(t_low,al$tl)
    t_high <- c(t_high,al$tu)
  }
  lower_prob <- sum(low, na.rm = TRUE)/ntt
  upper_prob <- sum(high,na.rm=TRUE)/ntt
  mean_tl <- mean(t_low)
  mean_up <- mean(t_high)
return(c(lower_prob,upper_prob,mean_tl,mean_up)) ## The outputs of the function are:
    lower RP estimate, bigger RP estimate mean of t_l's, mean of t_u's
}
```

NPI-B-RP for the GR inhibition significance (Algorithm 10)

```
reproducibility_GR_inhibition_finiteI <- function(dose1,dose2) {
  n <- length(dose1)
  avl <- vector()
  for (k in (1:(n-1))) {
    avl1 <- dose1[(k+1)] - dose1[k]
    avl <- c(avl,avl1)
  }
  avl_mean <- max(avl)
  n2 <- length(dose2)
  avl2 <- vector()
  for (k in (1:(n2-1))) {
    avl_2 <- dose2[(k+1)] - dose2[k]
    avl2 <- c(avl2,avl_2)
  }
  avl_mean2 <- max(avl2)
  so1 <- min(dose1) - avl_mean # min for dose1
  sn1 <- max(dose1) + avl_mean # max for dose1
  so2 <- min(dose1) - avl_mean2 # min for dose2
  sn2 <- max(dose1) + avl_mean2 # max for dose2
  one_times_step_finite_npi <- function(dose1,dose2,so1,sn1,so2,sn2) {
    function_comparison_finite_npi <- function(dose1,dose2,so,sn,so2,sn2) { # This
    function calculates $p$-value between two new bootstrap samples
      finite_npi <- function(x, so, sn) {
        m <- length(x)
        x <- append(x, c(so, sn), after =  length(x)) # Add starting and ending point
        x <- sort(x)
        boot <- vector()
        for (j in 1:m){ # This cycle creates m new values from the original intervals
          int_1 <- length(x) - 1
          jj<-sample(1:int_1,1,prob=rep(1/int_1,int_1))  # Sample an interval
          new_value <- runif(1, min = x[jj], max = x[jj+1]) # Sample a value in that
    interval
          x <- append(x, new_value, after = jj) # Add this to the set of values
```

```
        boot <- c(boot, new_value)
      }
      return(t(boot))
    }
    x1 <- finite_npi(dose1,so1,sn1)
    y1 <- finite_npi(dose2,so2,sn2)
    # GR inhibition = $(1 - \mu T/ \mu C)*100$ and of interest is 30\% inhibition
    GR=(1-mean(y1)/mean(x1))*100 >30
    wilc = wilcox.test(dose1,dose2, alternative = "greater", paired = FALSE)$p.value <
  0.05
    significantresult = GR == T & wilc == T
    return(significantresult)
  }
  total <- sum(replicate(1000, function_comparison_finite_npi(dose1,dose2,so1,sn1,so2,
  sn2)) == T)
  if ((1-mean(dose2)/mean(dose1))*100 > 30) {
    rp <- total/1000}
  else {rp <- 1 - total/1000}
  return(rp)
}
output <- replicate(100, one_times_step_finite_npi(dose1, dose2, so1,sn1,so2,sn2))
return(c(min(output),mean(output),max(output)))
}
```

Bibliography

[1] Aitkin, M. (2008). Applications of the bayesian bootstrap in finite population inference. *Journal of Official Statistics*, 24, 21–51.

[2] Alghamdi, F. M. (2022) Reproducibility of Statistical Inference Based on Randomised Response Data. *A thesis*, University of Durham. `http://etheses.dur.ac.uk/14783/1/PhD-Thesis-Fatimah-Alghamdi.pdf` [Accessed: 3 May 2023].

[3] Allison, D., Brown, A. W., George, B. J et al. (2016). Reproducibility: A tragedy of errors. *Nature*, 530, 27–29.

[4] Al Luhayb, A. S. M., Coolen, F. P. A. and Coolen-Maturi, T. (2023). Smoothed bootstrap for right-censored data. *Communications in Statistics – Theory and Methods*, DOI: 10.1080/03610926.2023.2171708.

[5] Alqifari, H. N. (2017). Nonparametric predictive inference for future order statistics. *A thesis*, University of Durham. `https://npi-statistics.com/pdfs/theses/HA17.pdf` [Accessed: 29 June 2022].

[6] Amrhein, V., Greenland, S. and McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, 567, 305–307.

[7] Anderson, J. A., Eijkholt, M. and Illes, J. (2014). Ethical reproducibility: towards transparent reporting in biomedical research. *Nature Methods*, 10, 843–845.

[8] Angelis, D. and Young, G. A. (1992). Smoothing the bootstrap. *International Statistical Review*, 60, 45–56.

[9] Appelbaum, M., Kline, R.B., Nezu, A. M., Cooper, H., Mayo-Wilson, E. and Rao, S. M. (2018). Journal Article Reporting Standards for Quantitative Research in

Psychology: The APA Publications and Communications Board Task Force Report. *American Psychologist*, 73, 3–25.

[10] ARRIVE (2023). ARRIVE guidelines. `https://arriveguidelines.org/arrive-guidelines` [Accessed: 8 August 2022].

[11] Pryseley, A., Mintiens, K., Knapen, K., Van der Stede, Y. and Molenberghs, G. (2010). Estimating precision, repeatability, and reproducibility from Gaussian and non-Gaussian data: A mixed models approach. *Journal of Applied Statistics*, 37, 1729–1747.

[12] Atmanspacher, H. and Maasen, S. (2016). *Reproducibility: Principles, problems, practices, and prospects*. Wiley, New Jersey.

[13] Augustin, T., Coolen, F. P. A., de Cooman, G. and Troffaes, M. C. M (2014). Introduction to Imprecise Probabilities. Wiley, Somerset.

[14] Bailey, D. H., Borwein, J. M. and Stodden, V. (2016). Facilitating reproducibility in scientific computing: principles and practice. In Atmanspacher, H. and Maasen, S. (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 115–140, Wiley, New Jersey.

[15] Baker, R. M., Coolen-Maturi, T. and Coolen, F. P. A. (2017). Nonparametric predictive inference for stock returns. *Journal of Applied Statistics*, 44, 1333–1349.

[16] Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533, 452–454.

[17] Baker, M. (2016). Quality time: it may not be sexy, but quality assurance is becoming a crucial part of lab life. *Nature*, 529, 456–458.

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

[19] Barba, L. A. (2018). Terminologies for reproducible research. George Washington University. `https://arxiv.org/pdf/1802.03311.pdf` [Accessed: 17 May 2022].

[20] Barber, J. A. and Thompson, S. G. (2000). Analysis of cost data in randomized trials: an application of the non-parametric bootstrap. *Statistics in Medicine*, 19, 3219–3236.

[21] Barnhart, H. X., Yow, E., Crowley, A. L. et al. (2016). Choice of agreement indices for assessing and improving measurement reproducibility in a core laboratory setting. *Statistical Methods in Medical Research*, 25, 2939–2958.

[22] Bayarri, M. J. and Mayoral, A. M. (2002). Bayesian design of "successful" replications. *The American Statistician*, 56, 207–214.

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

[24] Bendtsen, C., Kietzmann, M., Korn, R., Mozley, P. D., Schmidt, G. and Binnig G. (2011). X-ray computed tomography: semiautomated volumetric analysis of late-stage lung tumors as a basis for response assessments. *International Journal of Biomedical Imaging*, DOI: 10.1155/2011/361589.

[25] Benjamin, D. J., Berger, J. O., Johannesson, M. et al. (2018). Redefine statistical significance. *Nature Human Behaviour*, 2, 6–10.

[26] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, B*, 57, 289–300.

[27] Beran, R. Bootstrap asymptotics. *Encyclopedia of Mathematics*, `http://encyclopediaofmath.org/index.php?title=Bootstrap_asymptotics&oldid=37732` [Accessed 19 March 2021].

[28] Berger, J. (2012). Reproducibility of science: p-values and multiplicity. *Duke University, 8^th International Purdue Symposium on Statistics June 21, 2012*, `http://www.stat.purdue.edu/symp2012/docs/Purdue_Symposium_2012_Jim_Berger_Slides.pdf` [Accessed 7 May 2019].

[29] Bergman, R. G. and Danheiser, R. L. (2016). Reproducibility in Chemical Research. *Angewandte Chemie International Edition*, Editorial, 55, 12548–12549.

[30] Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, 73, 291–295.

[31] BinHimd, S. (2014). Nonparametric predictive methods for bootstrap and test reproducibility. *A thesis*, University of Durham. `https://npi-statistics.com/pdfs/theses/SB14.pdf` [Accessed: 29 June 2022].

[32] Blackman, N. J.-M. (2004). Reproducibility of clinical data I: continuous outcomes. *Pharmaceutical Statistics*, 3, 99–108.

[33] Blackman, N. J.-M. (2004). Reproducibility of clinical data II: categorical outcomes. *Pharmaceutical Statistics*, 3, 109–122.

[34] Bodden, C., von Kortzfleisch, V. T., Karwinkel, F., Kaiser, S., Sachser, N. and Richter, S. H. (2019). Heterogenising study samples across testing time improves reproducibility of behavioural data. *Scientific Reports*, 9.

[35] Bogomolov, M. and Heller, R. (2013). Discovering findings that replicate from a primary study of high dimension to a follow-up study. *Journal of the American Statistical Association*, 108, 1480–1492.

[36] Botvinik-Nezer, R. Wager, T. D. (2022). Reproducibility in neuroimaging analysis: challenges and solutions. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*.

[37] Bretz, F. and Westfall, P. H. (2014). Multiplicity and replicability: two sides of the same coin. *Pharmaceutical Statistics*, 13, 343–344.

[38] Buckheit, J. B. and Donoho, D. L. (1995). WaveLab and Reproducible Research. Technical report, Standford, CA.

[39] Buzney, E. A. and Kimball, A. B. (2008). A critical assessment of composite and coprimary endpoints: a complex problem. *Journal of the American Academy of Dermatology*, 59, 890–896.

[40] Carpenter, J. and Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine*, 19, 1141–1164.

[41] Chernick, M. R. (2007). *Bootstrap methods: A guide for practitioners and researchers*, 2nd Edition, Wiley.

[42] Chernick, M. R. and LaBudde, R. A. (2011). *An introduction to bootstrap methods with applications to R*. Somerset: Wiley.

[43] Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

[44] Collberg, C., Proebsting, T., Moraila, G., Shankaran, A., Shi, Z. and Warren, A. M. (2014). Measuring reproducibility in computer systems research. Technical report. Department of Computational Science, University Arizona, Tucson. `http://reproducibility.cs.arizona.edu/tr.pdf` [Accessed 27 May 2022].

[45] Collins, F., Tabak, L. (2014). Policy: NIH plans to enhance reproducibility. *Nature*, 505, 612–613.

[46] Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society open science*, 1:140216.

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

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

[49] Coolen, F. P. A. and Alqifari, H. N. (2017). Nonparametric predictive inference for reproducibility of two basic tests based on order statistics. *Journal of Statistical Theory and Practice*, 8, 591–618.

[50] Coolen, F. P. A., Coolen-Maturi, T. and Alqifari, H. N. (2018). Nonparametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, 47, 2527–2548.

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

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

[53] Coolen, F. P. A. and BinHimd, S. (2020). Nonparametric predictive inference bootstrap with application to reproducibility of the two-sample Kolmogorov-Smirnov test. *Journal of Statistical Theory and Practice*, 14:26.

[54] Coolen, F. P. A., Coolen-Maturi, T. and Al-nefaiee, A. H. (2014). Nonparametric predictive inference for system reliability using the survival signature. *Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, 228, 437–448.

[55] 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, 62.

[56] Coolen, F. P. A. and Yan, K. J. (2003). Comparing two groups of lifetime data. *In: ISIPTA 03: Proceedings of the Third International Symposium on Imprecise Probabilities and their Applications*, ed. Bernard, J., Seidenfeld, T. and Zaffalon, M., 148–161.

[57] Coolen, F. P. A. and Yan, K. J. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126, 25–54.

[58] Coolen-Maturi, T. and Elkhafifi, F. F. and Coolen, F. P. A. (2014). Three-group ROC analysis: A nonparametric predictive approach. *Computational Statistics & Data Analysis*, 78, 69–81.

[59] Crowley, P. H. (1992). Resampling methods for computation-intensive data analysis in ecology and evolution. *Annual Review of Ecology and Systematics*, 23, 405–447.

[60] Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, 3, 286–300.

[61] Cambridge English Dictionary (2022). data Meaning [online]. Available at: `http://dictionary.cambridge.org/dictionary/english/data` [Accessed 25 May 2022].

[62] Chung, M. K. (2007). The gaussian kernel. Lecture notes for *Medical Image Analysis*. `https://pages.stat.wisc.edu/~mchung/teaching/MIA/reading/diffusion.gaussian.kernel.pdf` [Accessed 7 December 2022].

[63] DasGupta, A. (2008). *Asymptotic theory of statistics and probability*, Springer, New York.

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

[65] De Capitani, L. (2013). An introduction to RP-Testing. *Epidemiology Biostatistics and Public Health*, 10, 1–16.

[66] De Capitani, L. and De Martini, D. (2013). Reproducibility probability estimation and testing for the Wilcoxon rank-sum test. *Journal of Statistics Computation and Simulation*, 85, 468–493.

[67] De Capitani, L. and De Martini, D. (2016). Reproducibility probability estimation and RP-testing for some nonparametric tests. *Entropy*, 18, 1–17.

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

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

[70] Dictionary (2023). conclusion Meaning [online]. Available at: `https://www.dictionary.com/browse/conclusion` [Accessed 14 February 2023].

[71] Dixon, P. M. (2006). Bootstrap Resampling. *Encyclopedia of Environmetrics*, El-Shaarawi, A. H. and Piegorsch, W.W. and Høst, G. (Eds.). https://doi.org/10.1002/9780470057339.vab028 [Accessed 15 May 2021].

[72] Donoho, D. L. (2010). An invitation to reproducible computational research. *Biostatistics*, 11, 385–388.

[73] Dwivedi, A. K. and Mallawaarachchib, I. and Alvarado, L. A. (2017). Analysis of small sample size studies using nonparametric bootstrap test with pooled resampling method. *Statistics in Medicine*, 36, 2187–2205.

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

[75] Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods. *Biometrika*, 68, 589–599.

[76] Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*, Philadelphia: Society for Industrial and Applied Mathematics.

[77] Efron, B. (1983). A leisurely look at the bootstrap, the jacknife, and cross-validation. *The American Statistician*, 37, 36–48.

[78] Efron, B. and Tibshirani, R. J. (1994). *An introduction to the bootstrap*, Chapman & Hall/CRC.

[79] Ehm, W. (2016). Reproducibility from the perspective of meta-analysis. In Atmanspacher, H. and Maasen, S. (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 141–167, Wiley, New Jersey.

[80] EQUATOR. Enhancing the QUAlity and Transparency Of health Research. https://www.equator-network.org/ [Accessed: 5 October 2023].

[81] Errington, T. M., Denis, A., Perfito, N., Iorns, E. and Nosek, B. A. (2021). Challenges for assessing replicability in preclinical cancer biology. *eLife*,10:e67995.

[82] Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E. and Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10: e71601.

[83] Falk, M. and Reiss, R. D. (1989). Weak convergence of smoothed and non-smoothed bootstrap quantile estimates. *The Annals of Probability*, 17, 363–371.

[84] Fenwick, N., Griffin, G. and Gauthier, C. (2009). The welfare of animals used in science: how the "Three Rs" ethic guides improvements. *Canadian Veterinary Journal*, 50, 523–530.

[85] Festing, M. F. W. (2014). Randomized block experimental designs can increase the power and reproducibility of laboratory animal experiments. *Institute for Laboratory Animal Research Journal*, 55, 472–476.

[86] Folkers, G. and Baier, S. (2016). A continuum of reproducible research in drug development. In Atmanspacher, H. and Maasen, S. (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 315–323, Wiley, New Jersey.

[87] Gamble, C., Krishan, A. and Stocken, D. (2017). Guidelines for the content of statistical analysis plans in clinical trials. *Journal of the American Medical Association*, 318, 2337–2343.

[88] Geisser, S. (1993). *Predictive inference: An introduction.* Chapman & Hall., New York.

[89] Gentleman, R. and Lang, D. T. (2007). Statistical analyses and reproducible research. *Journal of Computational and Graphical Statistics*, 16, 1–23.

[90] Gibb, B. C. (2014). Reproducibility. *Nature Chemistry*, 6, 653–654.

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

[92] Goodman, S. N., Fanelli, D. and Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8, DOI: 10.1126/scitranslmed.aaf5027.

[93] Gosselin R.-D. (2019). Guidelines on statistics for researchers using laboratory animals: the essentials. *Laboratory Animals*, 53, 28–42.

[94] Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379, 20200210.

[95] Haidich, A. B. (2010). Meta-analysis in medical research. *Hippokratia*, 14, 29–37.

[96] Hall, P. (1992). *Bootstrap and Edgeworth Expansion*, Springer, New York.

[97] Hall, P. and Wilson, S. R. (1991). Two guidelines for bootstrap hypothesis testing. *Biometrics*, 47, 757–762.

[98] Halsey, L. G. (2019). The reign of the p-value is over: what alternative analyses could we employ to fill the power vacuum? *Biology Letters*, 15, DOI:10.1098/rsbl.2019.0174.

[99] Halsey, L. G., Curran-Everett, D., Vowler, S. L. and Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature Methods*, 12, 179–185.

[100] Hather, G., Liu, R. et al. (2014). Growth rate analysis and efficient experimental design for tumor xenograft studies. *Cancer Informatics*, 13, 65–72.

[101] Held, L. (2020). A new standard for the analysis and design of replication studies. *Journal of the Royal Statistical Society: Series A*, 183, 431–448.

[102] Heller, R., Bogomolov, M. and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. *Proceedings of the National Academy of Sciences*, 111, 16262–16267.

[103] Heller, R. and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, 8, 481–498.

[104] Herrington, R. (2001). Simulating statistical power curves with the bootstrap and robust estimation. *A thesis*, University of North Texas.

[105] Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, 58, 295–300.

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

[107] Hill, B. M. (1988). De Finetti's theorem, induction, and A(n) or Bayesian nonparametric predictive inference (with discussion). *Bayesian Statistics 3*, Bernardo, J. M. et al. (Eds.). Oxford University Press, 211–241.

[108] Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods* (2nd edition). Wiley.

[109] Holmes, T. H. and He, X.-S. (2016). Human immunophenotyping via low-variance, low-bias, interpretive regression modeling of small, wide data sets: Application to aging and immune response to influenza vaccination. *Journal of Immunological Methods*, 437, 1–12.

[110] Hutson, A. D. and Ernst, M. D. (2000). The exact bootstrap mean and variance of an L-estimator. *Journal of the Royal Statistical Society B*, 62: 89–94.

[111] Hutson, A. D. (2002). A semi-parametric quantile function estimator for use in bootstrap estimation procedures. *Statistics and Computing*, 12, 331–338.

[112] Hutson A. D. (2022): The generalized sigmoidal quantile function. *Communications in Statistics – Simulation and Computation.*

[113] Hutton, H. N. and Williamson, P. R. (2000). Bias in meta-analysis due to outcome variable selection within studies. *Applied Statistics*, 49, 359–370.

[114] Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician*, 50, 361–365.

[115] ICTRP. International Clinical Trials Registry Platform. `https://www.who.int/clinical-trials-registry-platform` [Accessed: 13 October 2023].

[116] ISO (2017). Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation. https://www.iso.org/obp/ui/#iso:std:iso:21748:ed-2:v1:en [Accessed 2 November 2022].

[117] ISRCTN registry. International Standard Randomised Controlled Trial Number registry. https://www.isrctn.com/ [Accessed: 13 October 2023].

[118] Ioannidis, J. P. A. (2005). Why most published research findings are false. *Public Library of Science Medicine*, 2, e124.

[119] Ioannidis, J. P. A. (2014). How to make more published research true. *Public Library of Science Medicine*, 11, e1001747.

[120] Ioannidis, J. P. A. (2019). The importance of predefined rules and prespecified statistical analyses: Do Not Abandon Significance. *Journal of the American Medical Association*, 321, 2067–2068.

[121] Iorns, E., Gunn, W., Erath, J., Rodriguez, A., Zhou, J. et al. (2014). Replication attempt: "effect of BMAP-28 antimicrobial peptides on leishmania major promastigote and amastigote growth: role of leishmanolysin in parasite survival". *Public Library of Science ONE*, 9: e114614.

[122] Iqbal, S. A., Wallach, J. D., Khoury, M. J., Schully, S. D. and Ioannidis, J. P. A. (2016). Reproducible research practices and transparency across the biomedical literature. *Public Library of Science Biology*, 14: e1002333.

[123] Jagtap, R. S., Kale, M. M. and Gedam, V. K. (2021). Tail aligned composite quantile estimator for bootstrapping of high quantiles. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 7, 494–515.

[124] Jarvis, M. F. and Williams, M. (2016). Irreproducibility in preclinical biomedical research: Perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences*, 37, 290–302.

[125] JCGM (2012). International vocabulary of metrology – Basic and general concepts and associated terms (VIM). 3rd edition.

[126] Joint Committee for Guides in Metrology (2008). Evaluation of measurement data – Guide to the expression of uncertainty in measurement. https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6 [Accessed 1 June 2022].

[127] Jung, K., Lee, J., Gupta, V. and Cho, G. (2019). Comparison of bootstrap confidence interval methods for GSCA using a Monte Carlo simulation. *Frontiers in Psychology*, 10:2215.

[128] Kafkafi, N. and Golani, I. et al. (2017). Addressing reproducibility in singlelaboratory phenotyping experiments. *Nature Methods*, 14, 462–464.

[129] Kaplan, D. M. and Hofmann, L. (2019). High-order coverage of smoothed Bayesian bootstrap intervals for population quantiles. *Working Papers 19-14*, Department of Economics, University of Missouri, revised 19 Sep 2020.

[130] Karp, N. A. (2018). Reproducible preclinical research? Is embracing variability the answer? *Public Library of Science Biology*, 16, e2005413.

[131] Karp, N. A., Wilson, Z., Stalker, E., Mooney, L., Lazic, S. E., Zhang, B. and Hardaker, E. (2020). A multi-batch design to deliver robust estimates of efficacy and reduce animal use - a syngeneic tumour case study. *Scientific Reports*, 10:6178.

[132] Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16, 345–353.

[133] Lau, A. T. C. (2009). What are repeatability and reproducibility. *ASTM Standardisation News*.

[134] Lawless, J. F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, 92, 529–542.

[135] Lecoutre, B., Lecoutre, M.-P. and Poitevineau, J. (2010). Killeen's probability of replication and predictive probabilities: How to compute, use, and interpret them. *Psychological Methods*, 15, 158–171.

[136] Leek, J. T. and Peng, R. D. (2015). P values are just the tip of the iceberg. *Nature*, 520, 612.

[137] LePage, R. and Billard, L. (1992). *Exploring the limits of bootstrap*, John Wiley and Sons, Inc., New York.

[138] LeVeque, R. J., Mitchell, I. N. and Stodden, V. (2012). Reproducible research for scientific computing: Tools and strategies for changing the culture. *Journal Computing in Science & Engineering*, 14, 13–17.

[139] Li, Q., Brown, J. B., Huang, H. and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5, 1752–1779.

[140] Macnaughton, D. B. (2019). Two problematic promises on page one of the TAS Special Issue on Statistical Inference. MatStat Research Consulting Inc. `https://matstat.com/macnaughton2019a.pdf` [Accessed: 20 June 2022].

[141] Mager, H. and Göller, G. (1997). Resampling methods in sparse sampling situations in preclinical pharmacokinetic studies. *Journal of Pharmaceutical Sciences*, 87, 372–378.

[142] Mammen, E. (1992). *When does bootstrap work? Asymptotic results and simulations*, Springer-Verlag, New York.

[143] Maraun, M. and Gabriel, S. (2010). Killeen's (2005) p_{rep} coefficient: logical and mathematical problems. *Psychological Methods*, 15, 182–191.

[144] Marques, F. J., Coolen, F. P. A. and Coolen-Maturi, T. (2019). Introducing non-parametric predictive inference methods for reproducibility of likelihood ratio tests. *Journal of Statistical Theory and Practice*, 13, 15.

[145] Marques, F. J., Coolen, F. P. A. and Coolen-Maturi, T. (2019). Approximations for the likelihood ratio statistic for hypothesis testing between two beta distributions. *Journal of Statistical Theory and Practice*, 13, 17.

[146] Maturi, T. (2010). Nonparametric predictive inference for multiple comparisons. *A thesis*, University of Durham. `https://npi-statistics.com/pdfs/theses/TM10.pdf` [Accessed: 29 June 2022].

[147] McAlinden, C., Khadka, J. and Pesudovs, K. (2011). Statistical methods for conducting agreement (comparison of clinical tests) and precision (repeatability or reproducibility) studies in optometry and ophthalmology. *Ophthalmic & Physiological Optics*, 31, 330–338.

[148] Meeden, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics & Probability Letters*, 16, 103–109.

[149] Miller, J. (2009). What is the probability of replicating a statistically significant effect? *Psychonomic Bulletin & Review: Theoretical and review articles*, 16, 617–640.

[150] Munafò, M. R., Nosek, B. A., Bishop, D. V. M. et al. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, 0021.

[151] Nature (2018). Challenges in irreproducible research [Special Issue]. `https://www.nature.com/collections/prbfkwmwvz` [Accessed 21 July 2022].

[152] National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*, Washington, DC: The National Academies Press. `https://doi.org/10.17226/21915` [Accessed 28 April 2023].

[153] National Academies of Sciences, Engineering and Medicine. (2019). *Reproducibility and replicability in science*, Washington, D.C: The National Academies Press. `https://doi.org/10.17226/25303` [Accessed 28 April 2023].

[154] Nosek, B. A., Hardwicke, T. E., Moshontz, H. et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, 73, 719–48.

[155] NSF (2015). Social, behavioral, and economic sciences perspectives on robust and reliable science. *Report of the Subcommittee on Replicability in Science Advisory Committee to the National Science Foundation Directorate for Social, Behavioral, and Economic Sciences* `http://www.nsf.gov/sbe/AC_Materials/SBE_Robust_and_Reliable_Research_Report.pdf` [Accessed 5 June 2022].

[156] Nuzzo, R. (2014). Scientific method: Statistical errors. *Nature*, 506, 150–152.

[157] Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, `doi:10.1126/science.aac4716` [Accessed 20 May 2022].

[158] Patil, P., Peng, R. D. and Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11, 539–544.

[159] Patil, P., Peng, R. D., and Leek, J. (2016). Supplement for "A statistical definition for reproducibility and replicability". `https://www.biorxiv.org/content/biorxiv/suppl/2016/07/29/066803.DC1/066803-1.pdf` [Accessed 20 May 2022].

[160] Peers, I. S., South, M. C., Ceuppens, P. R., Bright, J. D. and Pilling, E. (2014). Can you trust your animal study data? *Nature Reviews Drug discovery*, 13, 560.

[161] Peng, R. D. (2009). Reproducible research and Biostatistics. *Biostatistics*, 3, 405–408.

[162] Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334, 1226–1227.

[163] Peng, R. D., Dominici F. and Zeger S. L. (2006). Reproducible epidemiologic research. *American Journal of Epidemiology*, 163, 783–789.

[164] Peng, X. and Peng, G. and Gonzales, C. (2005). Paper SP 05 Power analysis and sample size estimation using bootstrap. *Eli Lilly and Company*, Indianapolis, IN.

[165] Philtron, D., Lyu, Y., Li, Q. and Ghosh, D. (2018). Maximum rank reproducibility: a nonparametric approach to assessing reproducibility in replicate experiments. *Journal of the American Statistical Association*, 113, 1028–1039.

[166] Pillow, J. (2016). *Lecture 21: Bootstrap and Permutation Tests.* `http://pillowlab.princeton.edu/teaching/mathtools16/slides/lec21_Bootstrap.pdf` [Accessed: 9 November 2023].

[167] Polansky, A. M. (2000). Stabilizing bootstrap-t confidence intervals for small samples. *The Canadian Journal of Statistics*, 28, 501-516.

[168] Polansky, A. M. and Schucany, W. R. (1997). Kernel smoothing to improve bootstrap confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59: 821–838.

[169] Posavac, E. J. (2002). Using p-value to estimate the probability of a statistically significant replication. *Understanding Statistics*, 1, 101–112.

[170] power.boot: Conducting power analysis based on bootstrap. `https://rdrr.io/cran/bmem/man/power.boot.html` [Accessed 16 March 2021].

[171] Prinz, F., Schlange, T. and Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Drug discovery*, 10, 328–329.

[172] Pusztai, L., Hatzis, C. and Andre, F. (2013). Reproducibility of research and preclinical validation: problems and solutions. *Nature Reviews Clinical Oncology*, 10, 720–724.

[173] Reynolds, P. S. (2022). Between two stools: preclinical research, reproducibility, and statistical design of experiments. *BMC Research Notes*, 15, 73.

[174] Richter, S. H. (2017). Systematic heterogenization for better reproducibility in animal experimentation. *Laboratory Animals*, 46, 343–349.

[175] Richter, S. H., Garner, J. P. and Würbel, H. (2009). Environmental standardization: cure or cause of poor reproducibility in animal experiments? *Nature Methods*, 6, 257–261.

[176] Richter, S. H., Kunert, J. and Würbel, H. (2010). Systematic variation improves reproducibility of animal experiments. *Nature Methods*, 7, 167–168.

[177] Ristić-Djurović, J. L., Ćirković, S. et al. (2018). Analysis of methods commonly used in biomedicine for treatment versus control comparison of very small samples. *Computer Methods and Programs in Biomedicine*, 157, 153–162.

[178] RSPCA and LASA. (2015). Guiding Principles on Good Practice for Animal Welfare and Ethical Review Bodies. A report by the RSPCA Research Animals Department and LASA Education, Training and Ethics Section. Jennings, M. (ed.)

[179] Schaduangrat, N., Lampa, S., Simeon, S., Gleeson, M. P., Spjuth, O. and Nantasenamat, C. (2020). Towards reproducible computational drug discovery. *Journal of Cheminformatics*, 12, DOI: 10.1186/s13321-020-0408-x.

[180] Schwab, M., Karrenbach, M. and Claerbout, J. (2000). Making scientific computations reproducible. *Computing in Science & Engineering*, 2, 61–67.

[181] The Science Exchange Network. Validating key experimental results via independent replication. `http://validation.scienceexchange.com/#/` [Accessed: 8 August 2022].

[182] Senn, S. (2002). A comment on 'A comment on replication, p-values and evidence'. *Statistics in Medicine* (Letter to the editor), 21, 2437–2444.

[183] Serret-Larmande, A., Kaltman, J. R. and Avillach, P. (2022). Streamlining statistical reproducibility: NHLBI ORCHID clinical trial results reproduction. *Journal of the American Medical Informatics Association Open*, 5, ooac001.

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

[185] Shao, J. and Tu, D. (1995). *The jackknife and bootstrap.* (electronic resource). New York: Springer.

[186] Sheather, S. J. and Marron, J. S. (1990). Kernel quantile estimators. *Journal of the American Statistical Association*, 85, 410–416.

[187] Shenhav, L., Heller, R. and Benjamini, Y. (2015). Quantifying replicability in systematic reviews: the r-value. Cornell University. `https://arxiv.org/pdf/1502.00088.pdf` [Accessed: 28 April 2023].

[188] Shiffrin, R. and Chandramouli, S. (2016). Model selection, data distributions, and reproducibility. In Atmanspacher, H. and Maasen, S. (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 115–140, Wiley, New Jersey.

[189] Sidi, Y. and Harel, O. (2018). The treatment of incomplete data: Reporting, analysis, reproducibility, and replicability. *Social Science & Medicine*, 209, 169–173.

[190] Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* Chapman and Hall/CRC.

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

[192] Simkus, A., Coolen, F. P. A., Coolen-Maturi, T., Karp, N. A. and Bendtsen, C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research*, 31, 673–688.

[193] Spanagel, R. (2022). Ten points to improve reproducibility and translation of animal research. *Frontier in Behavioral Neuroscience*, 16: 869511.

[194] Stahel, W. A. (2016). Statistical issues in reproducibility. In Atmanspacher, H. and Maasen, S. (Eds.), *Reproducibility: Principles, Problems, Practices, and Prospects*, 87–114, Wiley, New Jersey.

[195] Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology. *Frontiers in Psychology*, 8, 1–6.

[196] Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2, 1–19.

[197] Storey, J. D. (2003). The positive false discovery rate: a Bayesian interpretation and the q-value. *The annals of statistics*, 31, 2013–2035.

[198] Teixeira da Silva, J. A. (2015) Negative results: negative perceptions limit their potential for increasing reproducibility. *Journal of Negative Results in BioMedicine*, 14:12.

[199] Tiwari, K., Kananathan, S., Roberts, M. G. et al. (2021). Reproducibility in systems biology modelling. *Molecular Systems Biology*, 17:e9982.

[200] Tsukamoto, M., Hatabu, A., Takahashi, Y., Matsuda, H., Okamoto, K., Yamashita, N. and Takagi, T. (2013). Statistical evaluation of single-photon emission computed tomography image using smoothed bootstrap method. *Biological and Pharmaceutical Bulletin*, 36, 417–424.

[201] Vexler, A., Hutson, A. D. and Xiwei, C. (2016). *Statistical testing strategies in the health sciences.* Boca Raton: Chapman & Hall.

[202] Voelkl, B., Altman, N. S., Forsman, A. et al. (2020). Reproducibility of animal research in light of biological variation. *Nature Reviews Neuroscience*, 16, e2003693.

[203] Voelkl, B., Vogt, L. and Sena, E. S. and Würbel, H. (2018). Reproducibility of pre-clinical animal research improves with heterogeneity of study samples. *Public Library of Science Biology*, 16, e2003693.

[204] Walters, S. J. and Campbell, M. J. (2004). The use of bootstrap methods for analysing health-related quality of life outcomes (particularly the SF-36). *Health and Quality of Life Outcomes*, 2.

[205] Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing.* Monographs on Statistics and Applied Probability 60. Chapman & Hall.

[206] Wasserstein, R. L., Schirm, A. L. and Lazar, N. A. (2019). Moving to a world beyond $p < 0.05$. *The American Statistician*, 73, 1–19.

[207] Wilkinson, L. and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54, 594–604.

[208] Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. et al. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

[209] Xie, Y. (2014). knitr: A comprehensive tool for reproducible research in R. In Stodden, V., Leisch, F. and Peng, R. D. (Eds.), *Implementing Reproducible Research.* CRC Press. `osf.io/s9tya/` [Accessed: 26 May 2022].

[210] Young, M. (2015). *Permutation Testing and Bootstrapping: Non?parametric Approaches to Statistical Testing and Estimation.* NASA Poster. `https://ntrs.nasa.gov/api/citations/20150001882/downloads/20150001882.pdf` [Accessed: 9 November 2023].

[211] Zwaan, R. A., Etz, A., Lucas, R. E. and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, e120, 1–61.