Nonparametric Predictive Methods for Bootstrap and Test Reproducibility

SULAFAH MOHAMMEDSALEH BINHIMD

Nonparametric Predictive Methods for Bootstrap and Test Reproducibility

Sulafah BinHimd

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
University of Durham
England

February 2014

Dedicated to

My Father

who has been a great source of motivation and support

My Mother

for all her unlimited love and prayers

My lovely nephews and nieces

for lighting up my life with their smiles

My Brothers and their wives

for all their love and best wishes throughout my life

All Family and Friends

for encouraging and believing

Nonparametric Predictive Methods for Bootstrap and Test Reproducibility

Sulafah BinHimd

Submitted for the degree of Doctor of Philosophy
February 2014

Abstract

This thesis investigates a new bootstrap method, this method is called Nonparametric Predictive Inference Bootstrap (NPI-B). Nonparametric predictive inference (NPI) is a frequentist statistics approach that makes few assumptions, enabled by using lower and upper probabilities to quantify uncertainty, and explicitly focuses on future observations. In the NPI-B method, we use a sample of n observations to create $n+1$ intervals and draw one future value uniformly from one interval. Then this value is added to the data and the process is repeated, now with $n+1$ observations. Repetition of this process leads to the NPI-B sample, which therefore is not taken from the actual sample, but consists of values in the whole range of possible observations, also going beyond the range of the actual sample. We explore NPI-B for data on finite intervals, real line and non negative observations, and compare it to other bootstrap methods via simulation studies which show that the NPI-B method works well as a prediction method.

The NPI method is presented for the reproducibility probability (RP) of some nonparametric tests. Recently, there has been substantial interest in the reproducibility probability, where not only its estimation but also its actual definition and interpretation are not uniquely determined in the classical frequentist statistics framework. The explicitly predictive nature of NPI provides a natural formulation of inferences on RP. It is used to derive lower and upper bounds of RP values (known as the NPI-RP method) but if we consider large sample sizes, the computation of

these bounds is difficult. We explore the NPI-B method to predict the RP values (they are called NPI-B-RP values) of some nonparametric tests. Reproducibility of tests is an important characteristic of the practical relevance of test outcomes.

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 all my own work unless referenced to the contrary in the text.

Acknowledgements

First, Allah my God, I am truly grateful for the countless blessing you have bestowed on me generally and in accomplishing this thesis especially.

My main appreciation and thanks goes to my supervisor Prof. Frank Coolen for his unlimited support, expert advice and guidance. It is very hard to find the words to express my gratitude and appreciation for him.

My special thanks go to my friends Tahani Maturi and Zakia Kalantan for all their support and help.

I would like also to thank my examiners Dr. Matthias Troffaes and Dr. Malcolm Farrow for their useful discussions.

Thanks to Durham University, King AbdulAziz University, Jeddah, Saudi Arabia, and Dr. Abeer for the facilities that have enabled me to study smoothly.

My final thanks go to everyone who has assisted me, stood by me or contributed to my educational progress in any way.

Sulafah Binhimd, Durham, UK
First Submit: September 2013
Viva: November 2013
Final Submit: February 2014

Contents

Abstract **iii**

Declaration **v**

Acknowledgements **vi**

1 Introduction **1**

1.1 Overview 1

1.2 Nonparametric Predictive Inference (NPI) 2

1.3 Bootstrapping 6

1.3.1 Efron's Bootstrap 6

1.3.2 Bayesian Bootstrap 11

1.3.3 Banks' Bootstrap 12

1.3.4 Comparison of Bootstrap Methods 13

1.4 Reproducibility 17

1.5 Outline of Thesis 22

2 NPI Bootstrap **24**

2.1 Introduction 24

2.2 The General Idea of NPI Bootstrap 25

2.3 NPI Bootstrap for Finite Intervals 28

2.4 NPI Bootstrap on Real Line 35

2.5 Comparison With Other Methods 43

2.5.1 Confidence Intervals 43

2.5.2 Prediction Intervals 46

2.6 NPI-B for Order Statistics 60
2.7 Concluding Remarks 62

3 NPI for Reproducibility of Basic Tests 64
3.1 Introduction 64
3.2 Overview of Some Basic Tests 65
3.2.1 One Sample Sign Test 65
3.2.2 One Sample Signed Rank Test 66
3.2.3 Two Sample Rank Sum Test 67
3.3 NPI for the Reproducibility Probability 69
3.4 NPI-RP for the One Sample Sign Test 72
3.5 NPI-RP for the One Sample Signed Rank Test 82
3.6 NPI-RP for the Two Sample Rank Sum Test 92
3.7 Concluding Remarks 101

4 Reproducibility using NPI-Bootstrap 104
4.1 Introduction 104
4.2 NPI-B-RP for the One Sample Sign Test 105
4.3 NPI-B-RP for the One Sample Signed Rank Test 113
4.4 NPI-B-RP for the Two Sample Rank Sum Test 116
4.5 NPI-B-RP for the Two Sample Kolmogorov Smirnov Test 120
4.6 Performance of NPI-B-RP 122
4.6.1 Mean Square Error with NPI-B-RP 122
4.6.2 Predictive Performance of NPI-B-RP 125
4.7 Concluding Remarks 128

5 Conclusions 129

Bibliography 132

Chapter 1

Introduction

1.1 Overview

Recently nonparametric predictive inference (NPI) [6, 17, 19] has been developed as a frequentist statistical approach that uses few assumptions. It uses lower and upper probabilities for events of interest considering future observations, and is based on Hill's assumption $A_{(n)}$ [45–47]. The lower and upper probabilities of NPI are introduced by Coolen and Augustin [6], they showed that NPI has strong consistency properties in the theory of interval probability [25, 68].

The standard bootstrap (standard-B) was introduced by Efron [35]. It is a resampling method for statistical inference, and a computer based method for assigning measures of accuracy to statistical estimates. Thereafter, various versions of bootstrap were developed such as smoothed Bayesian bootstrap, which was developed by Banks' [8].

In this thesis we present the NPI bootstrap method, which we indicate by NPI-B, as an alternative to other well known bootstrap methods, and then we discuss its performance, and use it to predict the reproducibility probability (RP). The reproducibility probability is a helpful indicator of the reliability of the results of statistical tests. The predictive nature of NPI provides a natural formulation of inference on reproducibility probability which is an important characteristic of statistical test outcomes. In this thesis we consider RP within a frequentist statistical framework but from the perspective of prediction instead of estimation.

Section 1.2 presents a brief introduction to NPI. In Section 1.3 we describe various bootstrap methods and Section 1.4 reviews the reproducibility probability for tests. The outline of this thesis is provided in Section 1.5.

1.2 Nonparametric Predictive Inference (NPI)

In classical probability, a single (precise) probability $P(A)$ is used for each event A but if the information is vague, the "imprecise probability" is an alternative approach which uses an interval probability instead of single probability $P(A)$. The interval probability is specified by lower and upper probabilities and denoted by $\underline{P}(A)$ and $\overline{P}(A)$, respectively, where $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, [25].

Nonparametric predictive inference (NPI) [6, 17, 19] is a statistical technique based on Hill's assumption $A_{(n)}$. Hill [45] introduced the assumption $A_{(n)}$ for prediction if there is no prior information about an underlying distribution. It is used to predict direct conditional probabilities for one future value Y_{n+1} or more than one future value $Y_i, i \geq n+1$. Conditional on the observed values, the $n+1$ intervals are created by n ordered, exchangeable and continuous random quantities on the real line $y_{(1)} < y_{(2)} < ... < y_{(n)}$, assigned equal probabilities for the next observation to belong to each of these intervals, which are denoted by $I_1 = (-\infty, y_{(1)})$, $I_l = (y_{(l-1)}, y_{(l)})$, for $l = 1, 2, ..., n+1$, and $I_{n+1} = (y_{(n)}, \infty)$. The assumption $A_{(n)}$ is:

$$P(Y_{n+1} \in I_l) = \frac{1}{n+1} \tag{1.1}$$

for $l = 1, 2, ..., n+1$. Hill gave more details about $A_{(n)}$ in [46] and [47]. It is clear that $A_{(n)}$ is a post data assumption and the statistical inferences based on it are predictive and nonparametric, and it is suitable if there is no knowledge about the random quantity of interest. $A_{(n)}$ is not sufficient to get precise probabilities for any event of interest, but it does give bounds (lower and upper) for probabilities which are called "interval (valued) probabilities" or "imprecise probabilities". They are lower and upper probabilities in interval probability theory [19, 68].

Coolen [15] gave an example to explain the dependence of Y related to $A_{(n)}$: Suppose we have a single observation y_1, providing two intervals, I_1, I_2. The assumption $A_{(1)}$ now states that $P(Y_i \in I_1) = P(Y_i < y_1) = \frac{1}{n+1} = \frac{1}{2}$ for all $i \geq 2$. Let us consider Y_3, and in particular how probability statements about Y_3 change when learning the value of Y_2. If we remain interested in the event $Y_3 < y_1$, the probability $P(Y_3 < y_1) = \frac{1}{2}$ will change, assuming $A_{(2)}$, according to whether the value of Y_2 will be less than or greater than y_1, $P(Y_3 < y_1 | Y_2 < y_1) = \frac{2}{3}$ or $P(Y_3 < y_1 | Y_2 > y_1) = \frac{1}{3}$, respectively. This is related to the probability $P(Y_3 < y_1) = \frac{1}{2}$ without conditioning on the unknown Y_2 by the theorem of total probability,

$$\begin{aligned} P(Y_3 < y_1) &= P(Y_3 < y_1 | Y_2 < y_1) P(Y_2 < y_1) + P(Y_3 < y_1 | Y_2 > y_1) P(Y_2 > y_1) \\ &= (\tfrac{2}{3} \times \tfrac{1}{2}) + (\tfrac{1}{3} \times \tfrac{1}{2}) = \tfrac{1}{2} \end{aligned}$$

A direct consequence of $A_{(2)}$ is that these probabilities for Y_3 keep the same values if the unknown Y_2 is replaced by its observed value y_2, so $P(Y_3 < y_1 | y_2 < y_1) = \frac{2}{3}$ and $P(Y_3 < y_1 | y_2 > y_1) = \frac{1}{3}$.

Augustin and Coolen [6] referred to the statistical approach known as nonparametric predictive inference (NPI) based on $A_{(n)}$. They introduced the lower and upper probabilities of NPI, and explained that NPI based only on the $A_{(n)}$ assumption has strong consistency properties in the theory of interval probability [25, 68]. NPI is exactly calibrated [53], which is a strong consistency property in the frequentist theory of statistics, and it never leads to results that are in conflict with inferences based on empirical probabilities.

The lower probability for an event A is denoted by $\underline{P}(A)$ and the upper probability by $\overline{P}(A)$. The lower probability in NPI is the maximum lower bound for the classical (precise) probability for A, $P(A)$, and the upper probability in NPI is the minimum upper bound for A, where $P(A) \in [0, 1]$. The classical (precise) probability is a special case of imprecise probability when $\underline{P}(A) = \overline{P}(A)$, $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, whereas $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ represent an absence of information about A. The NPI lower and upper probabilities for the event $Y_{n+1} \in B$ where $B \subset \mathbb{R}$ are :

$$\underline{P}(Y_{n+1} \in B) = \frac{1}{n+1} |\{l : I_l \subseteq B\}| \tag{1.2}$$

$$\overline{P}(Y_{n+1} \in B) = \frac{1}{n+1} |\{l : I_l \bigcap B \neq \emptyset\}| \tag{1.3}$$

The lower probability (1.2) is the total taking only probability mass into account that must be in B, which is only the case for the probability mass $\frac{1}{n+1}$ per interval I_l, if this interval is completely contained within B. The upper probability (1.3) is the total of all the probability mass into account that can be in B, which is the case for the probability mass $\frac{1}{n+1}$, per interval I_l, if the intersection of I_l and B is non empty.

NPI is used in a variety of statistical fields such as quality control [4, 5] and precedence testing [28]. The NPI method was explained in statistical process control by Arts, Coolen and van der Laan [5] who presented extrema charts and the run length distribution for these charts via simulation examples and compared them with other types of charts. Coolen and Coolen-Schrijner [21] used lower and upper probabilities for predictive comparison of different groups of proportions data. They considered NPI pair wise comparison of groups and then generalized to multiple comparisons. To illustrate their method and discuss the features, they analyzed two data sets and explained the importance of the choice of the number of future trials, then analyzed the imprecision in their results. Maturi et al [57] compared the failure times of units from different groups in life testing experiments if each unit failed at most once, and studied the effect of the early termination of the experiment on the lower and upper probabilities.

For data sets containing right-censored observations, Coolen and Yan [24] developed 'right censoring $A_{(n)}$' which is a generalization of $A_{(n)}$. Coolen-Schrijner and Coolen [27] used that generalization to find a method for the age replacement of technical units, and in simulations they found this method performed well. Coolen [17] explained NPI for circular data and multinomial data to show that it is possible to apply NPI to different data situations. He described two norms of objective Bayesianism, which both have a predictive nature.

The NPI approach is also used with multinomial data [7] when the observations fall into one of several unordered categories. It considers whether the number of categories is known or unknown. In the same paper the same idea is illustrated to deal with sub-categories, where it is assumed that the main category divides into several not overlapping sub-categories. In all cases, the multinomial data are

represented as observations on a probability wheel, like circular data. Recently presented NPI methods for statistical inference and decision support considered, for example, precedence testing [28], accuracy of diagnostic tests [26,39] and acceptance decisions [22,40]. In Chapter 3, we apply NPI to Bernoulli data [16] and to multiple real valued future observations [5].

NPI for Bernoulli random quantities [16] is based on a latent variable representation of Bernoulli data as real-valued outcomes of an experiment, in which there is a completely unknown threshold value, such that the outcomes on one side of the threshold are successes and on the other side are failures. The use of $A_{(n)}$ together with lower and upper probabilities enables inference without a prior distribution on the unobservable threshold value, as is needed in Bayesian statistics where this threshold value is typically represented by a parameter.

Assume that there is a sequence of $n+m$ exchangeable Bernoulli trials, that 'success' and 'failure' are possible outcomes of each one of these trials, and that the data consisting of s successes in n trials. Let Y_1^n denote the random number of successes in trials 1 to n, then a sufficient representation of the data for NPI is $Y_1^n = s$, due to the assumed exchangeability of all trials. Let Y_{n+1}^{n+m} denote the random number of successes in trials $n+1$ to $n+m$, or in future trials. Let $R_t = \{r_1, \ldots, r_t\}$, with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < \ldots < r_t \leq m$, and, for ease of notation, define $\binom{s+r_0}{s} = 0$. Then the NPI upper probability for the event $Y_{n+1}^{n+m} \in R_t$, given data $Y_1^n = s$, for $s \in \{0, \ldots, n\}$, is

$$\overline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = \binom{n+m}{n}^{-1} \sum_{j=1}^{t} \left[\binom{s+r_j}{s} - \binom{s+r_{j-1}}{s} \right] \binom{n-s+m-r_j}{n-s} \tag{1.4}$$

The corresponding NPI lower probability can be derived via the conjugacy property

$$\underline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \in R_t^c | Y_1^n = s) \tag{1.5}$$

where $R_t^c = \{0, 1, \ldots, m\} \backslash R_t$.

These NPI lower and upper probabilities are the maximum lower bound and minimum upper bound, respectively, for the probability for the given event based on the data, the assumption $A_{(n)}$ and the model presented by Coolen [16]. In Chapter 3 an explanation of the derivation of these NPI lower and upper probabilities is given, using a counting argument of paths on a grid. This is included in order to provide a combinatorial argument to prove an important claim in Section 3.4.

1.3 Bootstrapping

In this section we present a description of different bootstrap methods, some of them are used in this thesis, such as Efron's bootstrap and Banks' bootstrap.

1.3.1 Efron's Bootstrap

The bootstrap method was introduced by Efron [35]. It is a resampling technique for estimating the distribution of statistics based on independent observations, then developed to work with other statistical inferences. It is used for assigning the measures of accuracy of the sample estimate, especially the standard error. Using the bootstrap method to estimate the standard error does not require theoretical calculations, and it is available for any estimator. It is also a useful method when the sample size is not sufficient for statistical inference. The basic bootstrap method uses Monte Carlo sampling to generate an empirical estimate of the sampling distribution of the statistic (bootstrap distribution). That means it uses a plug-in principle to approximate the sampling distribution by using a bootstrap distribution. In most cases a bootstrap distribution mimics the shape, spread and bias of the actual sampling distribution. Monte Carlo sampling builds on drawing a large number of samples from the observations and finding the statistic for each sample. The relative frequency distribution of these statistics is an estimate of the sampling distribution of the statistic.

Efron [38] defined a bootstrap sample $x^* = (x_1^*, x_2^*, ..., x_n^*)$. It is obtained by randomly sampling n times with replacement, from the original sample $x_1, x_2, ..., x_n$. The size of a bootstrap sample can be chosen different to the size of the original sam-

ple. There are many references that show the principles and validity of bootstrap and how it works. Efron and Tibshirani [38] and Davison and Hinkley [31] have described bootstrap methods fully with examples and basic theories of applications, such as tests, confidence intervals and regression. Both books contain S-plus programs to implement this method. Efron and Gong [37] covered the nonparametric estimation of statistical errors, especially the bias and standard error of an estimator, in some resampling methods such as bootstrap, jackknife and cross validation. Efron [36] was concerned with the same basics but with more efficient computational methods for bootstrap and one sample problems.

Good [42] provided a brief review of bootstrap and computer code in various software packages (c++, SAS, MatLab, R, S-plus) in order to put this method into practice. Furthermore, Hjorth [48] explored FORTRAN code, but with more studies about bootstrap.

Chernick [13] discussed the key ideas and applications of bootstrap. Also he illustrated confidence intervals, hypothesis tests, regression models and time series. Singh [66] examined the convergence of the bootstrap approximation in some cases of estimation, and considered some of the theorems with proofs. Young [70] reviewed research into bootstrap and related methods, and additionally discussed the bootstrap of independent and dependent data.

We have the observations $x_1, x_2, ..., x_n$ of independent and identically distributed random variables $X_1, X_2, ..., X_n$ with distribution function F, and want to estimate the parameter of interest θ, which is a function of X, by statistic T_n. Now we would like to know the sampling distribution of the statistic T_n. To do this, we will use the bootstrap method, the main advantage of the bootstrap is that it can be applied to any statistic. There are two types of bootstrap: parametric and nonparametric bootstrap. The first type is the parametric bootstrap, which is used when we know the exact distribution function F (or the parametric form of the population distribution), or can assume it with some knowledge about the underlying population, and then estimate the parameters. The second type is the nonparametric bootstrap, which is used if F is completely unknown. This type is based on simulation of data from the empirical cumulative distribution function (CDF) F_n. F_n here is a discrete

probability distribution giving a probability $\frac{1}{n}$ to each value of the observations. The empirical distribution function is the maximum likelihood estimator of the population distribution function when there are no parametric assumptions. Here we show the basic steps of the nonparametric bootstrap, because it is the general method:

1. Construct F_n, the empirical probability distribution by putting probability $\frac{1}{n}$ to each value $x_1, x_2, ..., x_n$, $F_n(x) = \sum_{i=1}^{n} I(x_i \leq x)/n$. It is the number of elements which are less than or equal to x in the sample divided by size of this sample.

2. Draw B random samples of size n from F_n with replacement (from the original sample which is treated as a population).

3. Calculate the statistic of interest T_n from each sample to get $T_{n1}^*, T_{n2}^*, ..., T_{nB}^*$.

4. Construct the empirical distribution of $T_{n1}^*, T_{n2}^*, ..., T_{nB}^*$ by placing probability $\frac{1}{B}$ at each one of them. This distribution is a Monte Carlo approximation to the bootstrap estimate of the sampling distribution of the statistic T_n. It is used to make inferences about θ.

There are some possible modifications to this bootstrap procedure. Young [70] and Bickel and Freedman [10] explained a possibility of variation of the size of the data points and the size of the bootstrap sample.

The sampling distribution of a statistic shows the variation in the statistic, because the statistic will vary from one sample to another. If we use the bootstrap distribution as an approximation of a sampling distribution, we have another source of variation because we resample from the original sample. To solve this problem, we should use a large original sample and draw large numbers of bootstrap samples. To estimate the accuracy of an estimator T_n, the standard error of T_n, $se_F(T_n)$, is calculated. The bootstrap estimate of $se_F(T_n)$ is $se_{F_n}(T_n^*)$. It is a plug-in estimate because it uses the empirical distribution function F_n instead of the unknown distribution F. $se_{F_n}(T_n^*)$ is called the ideal bootstrap estimate of the standard error if $B \to \infty$, see [38], "idea" does not mean perfect. It simply refers to the use of an infinite number of bootstrap samples. To approximate $se_{F_n}(T_n^*)$, we use a Monte

Carlo approximation of the bootstrap estimate of the standard error $\widehat{se_B}$ by follow the next algorithm:

1. Draw B random samples of size n with replacement from the empirical distribution function F_n.

2. Calculate the statistic of interest T_n for each bootstrap sample to get $T^*_{n1}, T^*_{n2}, ..., T^*_{nB}$.

3. Estimate the standard error by the sample standard deviation of the T_{nj}, $j = 1, 2, ..., B$

$$\widehat{se_B} = \left[\frac{\sum_{j=1}^{B} \left(T^*_{nj} - T^*_n(.) \right)^2}{B-1} \right]^{0.5} \tag{1.6}$$

where $T^*_n(.) = \frac{\sum_{j=1}^{B} T^*_{nj}}{B}$.

Note that $\lim_{B\to\infty} \widehat{se_B} = se_{F_n}(T_n)$. So if B is very large, the difference between the bootstrap estimate and the Monte Carlo approximation will disappear. We use the bootstrap estimate of the standard error to compare between different bootstrap methods in Chapter 2.

The main advantage of the bootstrap method is that it enables us to estimate the standard error for any estimator. We discussed the standard error as a measure of accuracy for an estimator T_n, but there are other useful measures of statistical accuracy like bias, which is in general the difference between the expectation of an estimator T_n and the quantity θ being estimated,

$$bias(T_n, \theta) = bias_F = E(T_n) - \theta \tag{1.7}$$

Of course we want an estimator which has good characteristics such as small bias and small standard error. The bootstrap estimate of bias is

$$bias_{F_n} = E(T^*_n) - T^o_n \tag{1.8}$$

where T^o_n is the observed value of a statistic which is calculated from the original sample. Moreover, $bias_{F_n}$ is the ideal bootstrap estimate of bias. It is approximated by Monte Carlo simulation by generating independent bootstrap samples and evaluating the statistic T^*_n from each one, and approximating $E(T^*_n)$ by the average

$T_n^*(.) = \frac{\sum_{j=1}^{B} T_{nj}^*}{B}$ to get the bootstrap estimate of bias based on B replications:

$$\widehat{bias_B} = T_n^*(.) - T_n^o \tag{1.9}$$

If the ratio of bias to the standard error is small, we do not have to worry about bias. There is a better method than (1.9), see [38]. Let $P^* = (P_1^*, P_2^*, ..., P_n^*)$ be a resampling vector which contains the proportion of a bootstrap sample x^*

$$P_b^* = \frac{\#(x_i^* = x_b)}{n}, \quad b = 1, 2, ..., n \tag{1.10}$$

This vector satisfies $0 \leq P_b^* \leq 1$ and $\sum_{b=1}^{n} P_b^* = 1$. For example, if the bootstrap sample is $x^* = (x_1, x_6, x_6, x_5, x_1, x_1)$, then the corresponding resampling vector is $P^* = (3/6, 0, 0, 0, 1/6, 2/6)$. If the B bootstrap samples $x^{*1}, x^{*2}, ..., x^{*B}$ give B resampling vectors $P^{*1}, P^{*2}, ..., P^{*B}$, and $\overline{P}^*$ is the average of these vectors, then the better bootstrap bias estimate is

$$\overline{bias_B} = T_n^*(.) - T_n(\overline{P^*}) \tag{1.11}$$

where is $\overline{P^*} = \frac{\sum_{j=1}^{B} P_j^*}{B}$ and $T_n(\overline{P^*})$ is the bootstrap statistic but written as a function of the average of the resampling vector. Both $\overline{bias_B}$ and $\widehat{bias_B}$ converge to $bias_{F_n} = \widehat{bias}_\infty$, the ideal bootstrap estimate of bias, as $B \to \infty$, but the convergence is faster to $\overline{bias_B}$. Efron and Tibshirani [38] calculated, for some data, $\overline{bias_B}$ and $\widehat{bias_B}$ for $B = 25, 50, 100, ..., 3200$, and $\widehat{bias}_\infty$ approximated by $\widehat{bias}_{100,000}$, then they found $\overline{bias_B}$ converged to $\widehat{bias}_\infty$ more quickly.

The root mean square error of an estimator T_n for θ is a measure of accuracy that uses both bias and standard error [38]

$$\sqrt{MSE} = \sqrt{se(T_n)^2 + bias(T_n, \theta)^2} \tag{1.12}$$

Note that these procedures to find the bootstrap estimate of standard error and bias can be used in parametric and nonparametric bootstrap. In this thesis we use bias, variance, standard error and mean square error to compare different methods of bootstrap in Chapter 2.

Now, we should ask an important question, how large should we take B, the number of bootstrap replications? Efron [38] made some suggestions based on his experience. He considered that $B = 50$ replications is sufficient to estimate the standard error, but $B = 200$ replications or more is rarely needed to estimate a standard error. For confidence intervals and hypotheses tests $B = 50$ or 200 is not large enough. At least $B = 1000$ or 2000 replications is needed to give accurate results. Some researchers use even larger numbers but this consumes a lot of computer time, depending on the complexity of the method. In this thesis we use $B = 1000$ replications which we think is a suitable selection for our purposes here. It is the most widely used in the literature, especially for hypothesis tests and the construction of confidence intervals, and for the regression models.

1.3.2 Bayesian Bootstrap

The Bayesian bootstrap was developed by Rubin [63]. He explained that the Bayesian bootstrap simulates the posterior distribution of the parameter, whereas the standard bootstrap simulates the estimated sampling distribution of a statistic of the parameter. He used a noninformative prior distribution which is the uniform distribution. The standard and Bayesian bootstrap differ in how to assign the probabilities. In Bayesian bootstrap [13, 63], instead of sampling with replacement from the data and with probability $\frac{1}{n}$, it uses a posterior probability distribution for the data. To simulate the posterior distribution of the parameter Rubin [63] used the statistic as a function of probabilities g. For simplicity we will consider the data as one dimensional and as a single parameter but both can be multidimensional. He drew $n-1$ random variables from uniform (0,1) to get $u_1, u_2, ..., u_{n-1}$ and ordered them to have gaps g between these values. The vector g is the vector of probabilities of the data value $x_1, x_2, ..., x_n$. Then he drew a sample from the data and found the statistic of interest as a function of g. When he repeated this process the posterior distribution of the parameter was found. The Bayesian bootstrap can be used in the usual Bayesian inferences about the parameter θ, which is based on the estimated posterior distribution, but the nonparametric bootstrap just makes frequentist analysis about the distribution of statistic T_n.

Several studies have used and discussed the Bayesian bootstrap, such as Meeden [58], who presented Rubin's Bayesian bootstrap with a new modification. He used the same argument of the Bayesian bootstrap to estimate population quantiles but applied it to subintervals divided to grid, more than one grid is used, and these grids are given and fixed. Then Meeden compared the Bayesian bootstrap and the smoothed Bayesian bootstrap to his technique to show that these three methods are quite similar and preferable to traditional methods.

1.3.3 Banks' Bootstrap

Banks' [8] described new versions of the bootstrap, the smoothed Efron's bootstrap and the smoothed Bayesian bootstrap, and compared them to the Bayesian bootstrap and other bootstrap methods. Here we will focus on the smoothed Efron's bootstrap. In this method Banks' [8] smooths Efron's bootstrap by linear interpolation histospline smoothing between the jump points of empirical distribution. Histospline is a smooth density estimate based only on the information in a histogram. This procedure is as follows:

1. Take n observations, which are real valued, one dimensional on a finite interval.
2. Create $n+1$ intervals between the n observations $x_0, x_1, x_2, ..., x_n, x_{n+1}$ where x_0 and x_{n+1} are the end points of the possible data range (finite).
3. Put uniformly distributed probabilities $1/(n+1)$ over each interval.
4. Sample n observations from the distribution.
5. Find the statistic of interest.
6. Repeat steps 4 and 5 B times to get B bootstrap samples.

In smoothed Efron's bootstrap, the empirical distribution function $F_n(x)$ is smoothed using linear interpolation histospline smoothing between the jump points. It spreads the probability $1/(n+1)$ uniformly over any interval between two values of observations. Banks' [8] used confidence regions to compare his method to other bootstrap methods. Banks' estimated the confidence region at different values of α

and applied the chi-square test of goodness of fit to compare methods. The best region is one with small volume with α (Type I error).

1.3.4 Comparison of Bootstrap Methods

Efron [35,37,38] discussed different ways of comparing bootstrap methods like bias, standard error and mean square error, which are shown in Section 1.3.1. In this section, we illustrate other ways of comparison between bootstrap methods such as confidence intervals and prediction intervals.

1. Confidence Intervals

In this part we describe different methods for constructing confidence intervals by bootstrap technique. But we will start with a review of general confidence intervals. In one sample case, we have the observations $x_1, x_2, ..., x_n$ with the distribution function F, θ is the parameter of interest with its estimation T_n, $\widehat{se}$ is the estimate of the standard error of $T_n = \hat{\theta}$. In some cases, if the sample size n grows large, the distribution of T_n is approximated by normal with mean θ and variance $(\widehat{se^2})$, that means $T_n \sim N(\theta, (\widehat{se^2}))$ or equivalently

$$\frac{\hat{\theta} - \theta}{\widehat{se}} \sim N(0, 1) \tag{1.13}$$

This result is called the large sample theory or asymptotic theory.
Let $z^{(\alpha)}$ be the $100.\alpha$th percentile point of $N(0, 1)$, then from (1.13)

$$P(z^{(\alpha)} \leq \frac{\hat{\theta} - \theta}{\widehat{se}} \leq z^{(1-\alpha)}) = 1 - 2\alpha \tag{1.14}$$

and

$$P(\hat{\theta} - z^{(1-\alpha)}.\widehat{se} < \theta < \hat{\theta} - z^{(\alpha)}.\widehat{se}) = 1 - 2\alpha \tag{1.15}$$

It is called the standard confidence interval with coverage probability $1 - 2\alpha$, or confidence level $100.(1 - 2\alpha)\%$.

Now we want to explore the use of the bootstrap method to construct confidence intervals [38], the first approach is the bootstrap-t interval. We generate B bootstrap samples and for each one we find

$$Z^*(b) = \frac{\hat{\theta}^*(b) - \hat{\theta}}{\widehat{se^*_B(b)}} = \frac{T^*_{nj} - T_n}{\widehat{se^*_B(b)}} \tag{1.16}$$

where $\hat{\theta}^*(b)$ is the value of $\hat{\theta}$ of the bootstrap sample x^{*b} and $\widehat{se^*_B(b)}$ is the estimated standard error of $\hat{\theta}^*$ of the bootstrap sample x^{*b}. The bootstrap-t confidence interval is:

$$(\hat{\theta} - \hat{t}^{(1-\alpha)}.\widehat{se_B}, \hat{\theta} - \hat{t}^{(\alpha)}.\widehat{se_B}) \tag{1.17}$$

where $\hat{t}^{(\alpha)}$ is the αth percentile of $Z^*(b)$, across all bootstrap samples b, and $\hat{t}^{(1-\alpha)}$ is the $(1-\alpha)$th percentile. If $B = 1000$ and $\alpha = 0.05$, then $\hat{t}^{(\alpha)}$ is the 50th largest value of the $Z^*(b)$. If $B.\alpha$ is not an integer, assuming $\alpha \leq 0.5$, let $k = [(B+1)\alpha]$ is the largest integer$\leq (B+1)\alpha$, then determine α and $(1-\alpha)$ by the kth largest and $(B+1-k)$th largest value of $Z^*(b)$, respectively.

Another approach of confidence intervals using the bootstrap technique is based on the percentiles of the bootstrap distribution of a statistic. It is called the percentile interval. The approximate $(1-2\alpha)$ percentile interval is defined by :

$$\hat{\theta}^{*(\alpha)}_B < \theta < \hat{\theta}^{*(1-\alpha)}_B \tag{1.18}$$

$$T^{*(\alpha)}_{nj} < \theta < T^{*(1-\alpha)}_{nj} \tag{1.19}$$

where $T^{*(\alpha)}_{nj}$ is the $100.\alpha$th empirical percentile of the T^*_{nj} values and $T^{*(1-\alpha)}_{nj}$ is the $100.(1-\alpha)$th empirical percentile of them, that means the $B.\alpha$th value and $B.(1-\alpha)$th value of the ordered list of the B replications of T^*_n. For example, if $B = 1000$ and $\alpha = 0.05$, $T^{*(\alpha)}_{nj}$ and $T^{*(1-\alpha)}_{nj}$ are the 50th and 950th ordered values of the replications, respectively.

The percentile interval can be improved to get a BCa interval, or a bias corrected and accelerated interval. To find the endpoints which are given by percentiles, we need to compute two numbers $\hat{a}$ and $\hat{z}_0$, they are called the acceleration and bias correction, respectively. The value of the bias correction $\hat{z}_0$ uses the ratio of bootstrap replications less than the original estimate T_n, it counts the possible bias in T_n as an estimate of θ

$$\hat{z}_0 = \Phi^{-1}(\frac{\#(T^*_{nj} < T_n)}{B}) \tag{1.20}$$

where Φ^{-1} is the inverse function of a standard normal cumulative function, for example, $\Phi^{-1}(0.95) = 1.645$. To find the acceleration $\hat{a}$, we will use the easiest way

by jackknife values of a statistic T_n. Let $x_i = (x_1, x_2, ..., x_{i-1}, x_{i+1}, ..., x_n)$ be the jackknife sample which is the original sample with the ith observation x_i deleted, $T_{n(i)}$ is the $i-th$ jackknife replication of T_n and

$$T_n(.) = \frac{\sum_{i=1}^{n} T_{n(i)}}{n} \tag{1.21}$$

then

$$\hat{a} = \frac{\sum_{i=1}^{n}(T_{n(.)} - T_{n(i)})^3}{6(\sum_{i=1}^{n}(T_{n(.)} - T_{n(i)})^2)^{\frac{3}{2}}} \tag{1.22}$$

It refers to the rate of change of the standard error of T_n as θ varies.
The $(1-2\alpha)$ BCa interval in [38] and [56] is:

$$T_{nj}^{*(\alpha 1)} < \theta < T_{nj}^{*(\alpha 2)} \tag{1.23}$$

where

$$\alpha 1 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{\alpha}}{1 - \hat{a}(\hat{z}_0 + z^{\alpha})}) \tag{1.24}$$

$$\alpha 2 = \Phi(\hat{z}_0 + \frac{\hat{z}_0 + z^{1-\alpha}}{1 - \hat{a}(\hat{z}_0 + z^{1-\alpha})}) \tag{1.25}$$

$\Phi(.)$ is the standard normal cumulative function and $z^{(\alpha)}$ is the $100.\alpha$th percentile point of standard normal distribution. For example, $z^{(0.95)} = 1.645$ and $\Phi(1.645) = 0.95$. If $\hat{a}$ and $\hat{z}_0$ equal zero, then $\alpha 1 = \Phi(z^{(\alpha)}) = \alpha$, $\alpha 2 = \Phi(z^{(1-\alpha)}) = 1 - \alpha$, and the BCa interval and percentile interval are equal in this case.

We will use the BCa interval in Section 2.5.1 to compare different bootstrap methods because it has a higher order of accuracy and transformation respecting. To illustrate the transformation respecting property we consider an example, if we constructed a confidence interval for a parameter θ, then the interval for θ^2 will construct by squares the end points of confidence interval for θ, this interval that is transformation respecting [38].

2.Prediction Intervals

In [55], the bootstrap method was used to construct a prediction interval for one or more future values from the Birnbaum-Saunders distribution. They applied the bootstrap percentile method with the bootstrap calibration for estimating the prediction interval. They assumed that a random sample $x_1, ..., x_n$ is drawn from the

Birnbaum Saunders distribution function F with parameters α, β. A bootstrap sample of size n, $x_1^*, ..., x_n^*$, is drawn from $x_1, ..., x_n$ with replacement to construct the estimated distribution F^*, and then sampled $y_1^*, ..., y_m^*$ from it, (m is the number of future observations), we can obtain the mean of $y_1^*, ..., y_m^*$, which is denoted by $\bar{y}_m^*$. Repeat this technique B times to get B values of $\bar{y}_m^*$, denoted by $\bar{y}_m^*(1), ..., \bar{y}_m^*(B)$. The $1-\alpha$ prediction interval for $\bar{x}_m$ (the mean of future observations in the population) is: $[\bar{y}_{m,B}^{*(\frac{\alpha}{2})}, \bar{y}_{m,B}^{*(1-\frac{\alpha}{2})}]$. The lower bound $\bar{y}_{m,B}^{*(\frac{\alpha}{2})}$ is the $B.\frac{\alpha}{2}$th value in the ordered list of the B replications of $\bar{y}_m^*$, and the upper bound $\bar{y}_{m,B}^{*(1-\frac{\alpha}{2})}$ is the $B.(1-\frac{\alpha}{2})$th value in the same ordered list. To find the coverage count how many intervals contain $\bar{x}_m$. This procedure is used to predict one or more future observations. To explain this, for example, if we have a sample of x_1, x_2, x_3 , $n = 3$, and want to predict $m = 3$ values, x_4, x_5, x_6, the mean of these values is $\bar{x}_m = \bar{x}_3$. Then we construct the prediction interval for it by sampling the bootstrap sample x_1^*, x_2^*, x_3^* and then generating from them y_1^*, y_2^*, y_3^* and finding $\bar{y}_m^*$. Repeat this $B = 1000$ times and find the prediction interval of these B values as described before. If we consider the case of prediction of one future value x_4 the mean will not be used here. Simply resample x_1^*, x_2^*, x_3^* and generate y_1^* from them, and repeat this B times to have the list of B values $y_1^*(1), y_1^*(2), ..., y_1^*(B)$ to construct the prediction interval for $x_{n+1} = x_4$. In that study, the 90% and 95% prediction intervals for a single future value x_{n+1} and the mean of m future observations $\bar{x}_m$ are obtained, and MonteCarlo simulation is used to estimate the coverage probability by finding the proportion of intervals which contain x_{n+1} and $\bar{x}_m$.

Different types of bootstrap prediction intervals [2,60,61] can be used to estimate the parameter θ: bootstrap-t, percentile and BCa prediction intervals. In this thesis we use the percentile prediction interval. Let $X = (X_1, X_2, ..., X_n)$ be the past random samples and $Y_1, Y_2, ..., Y_m$ be the future random samples, where X and Y are iid from probability distribution F and $T = \hat{\theta}$ is a scalar parameter, and we want to construct the prediction interval to predict the statistic (estimator) of $\widehat{\theta_m} = T_m$ of the future random sample. Let $\widehat{\theta_n} = T_n$ be the estimator using the past sample of size n. F_n and F_m are the CDF of T_n and T_m, respectively, and let $\hat{F}_n$ and $\hat{F}_m$ be the CDF's of T_n^* and T_m^*. Here $\hat{F}_n$ assigns mass $\frac{1}{n}$ to each X_i^* and $\hat{F}_m$ assigns mass

$\frac{1}{m}$ on each Y_i^*. The $(1-2\alpha)\%$ percentile prediction interval

$$\text{lower bound} = \hat{F}_m^{-1}[\Phi(z^{(\alpha)}(1+\frac{m}{n})^{\frac{1}{2}})] = \hat{F}_m^{-1}[\alpha_1] \tag{1.26}$$

here, $r = \frac{m}{n}$ and $z^{(\alpha)} = \Phi^{-1}(\alpha)$

$$\text{upper bound} = \hat{F}_m^{-1}[\Phi(z^{(1-\alpha)}(1+\frac{m}{n})^{\frac{1}{2}})] = \hat{F}_m^{-1}[\alpha_1] \tag{1.27}$$

In [61] X^* and Y^* are drawn from the past sample X with replacement while [60] the iterated (calibration) bootstrap was used by generating X^* and Y^* from X and then resample Y^{**} from X^*, to study the improving of the coverage accuracy of prediction intervals. The percentile prediction interval to predict future observations, and the percentile prediction interval to predict the statistic, are used in Section 2.5.2 to compare the NPI-B method with the standard-B method.

1.4 Reproducibility

Often when we use the applications of the statistical test in several fields, we meet some problems because the results and conclusions of statistical hypothesis tests can be different each time the tests are repeated. Goodman [43] raised the topic of reproducibility of a statistical test, mainly to counter a frequently occurring misunderstanding about the meaning of a statistical p-value. The reproducibility probability (RP) for a test is the probability for the event that, if the test is repeated based on an experiment performed in the same way as the original experiment, the test outcome, that is either rejection of the null-hypothesis or not, will be the same. The focus is mostly on reproducibility of tests in which the null-hypothesis is rejected, as significant effects tend to lead to new treatments in medical applications, for example. In a later discussion of Goodman's paper (and about twice the length of Goodman's paper), Senn [64] emphasized the different nature of the p-value and RP. Senn agrees with Goodman about the importance of reproducibility of test results and the RP, but disagrees with Goodman's claim that 'p-values overstate the evidence against the null-hypothesis'. Indeed, it is important to recognize the difference between RP and the p-value, while also recognizing a natural link between the two, in the sense that the p-value indicates the strength of the statistical conclusion and the smaller

p-value in the case of a rejected null-hypothesis, the larger one would expect the corresponding RP to be.

Senn [64] also discusses the importance of reproducibility of tests in the real world, where actual repeats of tests may well be under slightly different circumstances and might involve different teams of analysts performing the tests. So, the concept of reproducibility is not necessarily uniquely defined, and statistical methodology should be flexible enough to deal with varying circumstances. Recently, Begley and Ellis [9] presented a worrying insight into problems with reproducing tests in preclinical cancer research in which significant effects were reported. Attempting to repeat tests from 53 'landmark studies', they managed to get the same significant scientific findings in only 6 cases. They report further on similar, but more extensive, studies by Bayer Healthcare in Germany, where only about 25 percent of significant results had been reproduced. Begley and Ellis [9] provide a detailed discussion of factors that play a role in such studies of repeatability and provide guidelines for improving the reliability of such studies which, for example, considers publication processes (there is an obvious bias due to the tendency for only 'positive' results to be published). Remarkably, Begley and Ellis [9] do not discuss the statistical methods used in such medical testing, where more emphasis on RP seems a natural requirement as part of a solution for more reliable medical testing.

During the last decade, the concept of RP has attracted increasing interest in the literature, with some contributions making clear that the definition and interpretation of RP are not uniquely determined. Miller [59] emphasizes that two scenarios for repetition of a test must be distinguished: a general form of repetition by other researchers, where conditions may vary with regard to the original experiment and test, and an individual form of repetition by the same researcher under exactly the same conditions as the original experiment and test. Miller [59] is sceptical about the possibility to derive useful inferences from one initial experiment, in particular as real effect sizes will be unknown, and hence the power of the test is unknown. The difference between these two scenarios is important, and links nicely to the discussions by Senn [64] and Begley and Ellis [9]. The approach to inference on RP presented in this thesis sits fully in the 'individual form of repetition' in Miller's

terminology, and makes clear that meaningful frequentist inference is possible in this scenario.

The recent literature on RP is fascinating, as it is clear that RP is not a natural concept in classical frequentist statistics. Shao and Chow [65] present three approaches to RP: a Bayesian approach, a frequentist approach based on estimating the power of a future test based on the available test data, and a corresponding approach where RP is linked to a lower confidence bound of this power estimate. While the Bayesian approach provides a natural solution to inference for the RP, through the use of the posterior predictive distribution, the unavoidable influence of the prior distribution [17] can lead to criticisms in medical scenarios where objectivity is essential. The natural way of formulating inference on RP as a predictive problem is also followed in the approach presented in this thesis, yet the statistical methodology presented here is fully frequentist. Shao and Chow [65] emphasize the possible use of RP in circumstances where evidence in a medical trial is overwhelmingly strong in favour of a new treatment. Currently, marketing approval of a new drug in the USA requires substantial evidence of its effectiveness in at least two clinical trials, although under special circumstances exceptions appear to be possible. Shao and Chow used several study designs to evaluate RP such as: two samples with equal variances, two samples with unequal variances and parallel group designs, and then they used them to study the generalization of the clinical results from one patient population to a different patient population, and also to adjust the sample size for the second trial.

With regard to estimation of RP, there is a simple argument that, if the distribution under the null-hypothesis of the test statistic is (about) symmetric, then a worst-case scenario would give RP of (about) 0.5 [43, 64]. This follows from the possibility that a value of the original test statistic could be equal to the critical value of the test. Without further information, one could expect that a repetition of the experiment would lead to a second value of the test statistic which is equally likely to be larger than or smaller than the original value, and therefore the same conclusion would be reached with probability 0.5 (Goodman [43] supports this with a Bayesian argument with a non-informative prior). Of course, it is more realistic to

consider the estimation problem in more detail, and not to be focused solely on the worst-case scenario of a test statistic that is equal to the critical value of the test.

It seems logical that RP should be considered as a function of the value of the original test statistic. The estimated power approach by Shao and Chow [65] uses the original test results to estimate the power of the test, assuming implicitly that the test would be repeated on the same number of units also sampled from the same probability distribution. This estimated power of the test is currently actually called the 'reproducibility probability' by several authors (e.g. De Martini [34]), and while it is a natural concept to consider in the classical frequentist statistics framework, it does not seem to be fully in line with the intuitive meaning of RP due to the explicit conditioning, for the power of a test, on the alternative hypothesis being true. The strength of support for this assumed condition in the results of the actual test depends again on the p-value, underlining the complications in defining RP within the classical frequentist framework.

Whether or not this concept for estimation of RP by Shao and Chow [65] is fully in line with the intuitive concept of RP, the approach has led to insightful further developments, particularly by De Martini [34] who considers such estimation with main focus on testing with one-sided alternative hypotheses (the theory for two-sided alternative hypotheses is also presented in an appendix). Importantly, De Martini [34] proposes not only to study such RP estimation for tests, but also to actually use the estimated RP to define tests, which provides an interesting alternative to tests based mainly on chosen statistically significant levels. De Martini managed several definitions of the RP of a statistic significant result. The first one is the power π_α of the test, and the second is the lower confidence bound of the power. This approach is followed by De Capitani and De Martini [32, 33] in detailed studies for the Wilcoxon Rank Sum (WRS) test, which we will also consider in this thesis. De Capitani and De Martini [32, 33] evaluated different RP estimators for the Wilcoxon rank sum WRS test and compared the performance of these estimators. Goodman [43] illustrated that the p-value gives too optimistic evaluation so De Capitani and De Martini [32] think it is suitable to use the RP estimate also. Collings and Hamilton [14] described the approximation of the power

of the Wilcoxon two sample test for the location shift model, but by bootstrap method without assumptions and if the shape of distribution is unknown.

Without attempting to provide a full review of the literature on this fascinating topic, it is worth to mention briefly some further contributions. Posavac [62] uses the difference between the value of a test statistic based on the actual test and the corresponding critical value to estimate RP. For example, for a two-sample test this requires the estimation of the standard error of the difference of the means of the two samples. This leads to an estimate of the probability of a 'statistically significant exact replication', which we believe is one of several examples of rather confusing concepts in this topic area.

Killen [51] emphasizes the predictive nature of RP and links it to the effect size. He proposes to effectively eliminate the effect size by averaging it out in a Bayesian manner with the use of a flat prior distribution. The paper is accompanied by a discussion which makes clear that the concept of RP is somewhat confusing, but the general ideas about RP in Killen's paper are not too distant from those presented in this thesis, namely the predictive nature of RP, which we will explicitly use, and the informal way of considering RP as the 'real power' of a test, with 'power' interpreted in its every-day, so non-statistical, meaning, which we also support. In [51] Killen defined the statistic p_{rep} as the estimate of RP. Consider an experiment test where there is no difference between experimental and control groups, so the null hypothesis is $H_0 : \mu_E - \mu_C = 0$. The observed effect size is $\acute{d}$ and the population effect size is $\acute{\delta}$

$$\acute{d} = \frac{M_E - M_C}{s_p} \tag{1.28}$$

where s_p is the pooled within group standard deviation, M_E and M_C are the sample means of the experimental group and control group, respectively. In [50] was considered the statistic p_{rep} as

$$p_{rep} = \Phi[\Phi^{-1}(1 - \frac{p}{2})/\sqrt{2}] \tag{1.29}$$

where Φ is the standard normal cumulative distribution function.

Lecoutre et al [54] discuss Killen's approach further, referring to it as 'fiducial Bayesian predictive probability', and mentioning that it is now increasingly popular. They discuss some problems with its computation, resulting again from some apparent confusions. They particularly emphasize the importance of predictive inference, ending their paper with 'Predictive probabilities are an unavoidable part of statistical thinking, and the time has come to take them seriously'- we wholeheartedly agree with this. Recently, Boos and Stefanski [12] considered reproducibility issues by studying the variability in p-values through bootstrap studies, which showed 'surprisingly large variability'. They also comment briefly on the importance of this issue in case of multiple testing, as increasingly used with very large data sets in, e.g., modern bio-statistics.

Cumming [29, 30] provided illustrations of Killen's statistics [51], but he considered that Killen's statistic is the average of all possible replication probabilities (or RP values). He described three ways to picture the idea of replication: confidence intervals, p-value and Killen's statistic.

To summarize, it is quite surprising that there is apparently confusion about reproducibility, which itself appears to be quite a straightforward concept. In this thesis, we consider RP within a frequentist statistical framework but from the perspective of prediction instead of estimation, which we think is attractive and avoids some of the confusion in earlier contributions.

1.5 Outline of Thesis

The work in this thesis proposes a new version of bootstrap, which is called non-parametric predictive inference bootstrap (NPI-B), and uses it to predict the reproducibility probability (RP). Also the NPI-RP is presented for the reproducibility probability.

In Chapter 2, we present the main idea of NPI-B and the difference between standard, Banks' and NPI bootstrap methods. NPI-B with finite intervals and real line observations is derived. A comparison of the three methods of bootstrap is pre-

sented using some measures like bias and mean squared error, then using confidence intervals and prediction intervals. Some results of this chapter were presented at the 1st International Statistical Conference with special sessions on Science, Engineering and Islamic Finance (ISM-1 2012) in Malaysia and they were also presented in the paper "On bootstrapping using nonparametric predictive inference" which is published in the proceedings of this conference [11].

Chapter 3 introduces a summary of three basic nonparametric tests, one sample sign test, one sample signed rank test and two sample rank sum test. This is followed by a demonstration of the use of NPI for the reproducibility probability (NPI-RP) of test results and includes some examples. We found the NPI lower and upper bounds of RP for various tests, with different sample sizes and levels of significance. The results of this chapter are included in the paper "Nonparametric predictive inference for reproducibility of basic nonparametric test", which has been accepted for publication in the Journal of Statistical Theory and Practice [20]. But the calculations of NPI-RP become complicated with large sample sizes or complex tests, for this reason we introduce an alternative method in Chapter 4.

Chapter 4 shows the reproducibility probability using the NPI bootstrap (NPI-B-RP). It also explores the reasons for using the NPI-B method rather than the NPI-RP method, and how the results in this chapter support those in Chapter 3. Then we show that NPI-B-RP can be applied to other tests by considering the Kolmogorov Smirnov test. At the end of this chapter we explore the performance of NPI-B to find RP. A paper presenting these results is in preparation.

In Chapter 5, we end with some remarks and conclusions. The calculations in this thesis were performed using the statistical software R version 2.12.2.

Chapter 2

NPI Bootstrap

2.1 Introduction

In this chapter we introduce the main concept of the NPI bootstrap (NPI-B), and how to derive an NPI-B sample for observations on finite and infinite intervals. The procedure depends on creating $n+1$ intervals using n observations, then drawing one value from these intervals and adding this value to the data set, and continuing to sample n further values in the same way in order to obtain an NPI-B sample. The assumptions are different with finite and infinite intervals. Additionally in this chapter different methods of bootstrap are compared, using confidence intervals and prediction intervals to discuss the strength of estimation and prediction inference of NPI-B, respectively.

Section 2.2 presents the main idea of NPI-B and explains the difference between standard, Banks' and NPI bootstrap methods. Section 2.3 explains how to derive an NPI-B sample from distributions which have restricted intervals, and simulation studies of this method are shown. In Section 2.4 we present the NPI-B approach with real line quantities and non negative observations using further assumptions. In Section 2.5 we compared different methods of bootstrap using confidence intervals and prediction intervals. Section 2.6 shows how NPI-B works with order statistics. Section 2.7 presents some concluding remarks.

2.2 The General Idea of NPI Bootstrap

In this section we present the main idea of the three types of bootstrap: standard, Banks' and NPI-B, and explain the difference between them. For the standard bootstrap method, the observations are drawn from the n original sample points, but with the other two kinds it is drawn from the points of the original sample and from the intervals between them. NPI-B depends on creating $n+1$ intervals using n observations, then drawing one value from these intervals and adding this value to the data set, and continuing to sample m further values in the same way in order to derive an NPI-B sample. Banks' bootstrap uses the same process but without adding the new value to the data set. The style of sampling observations of NPI-B, which samples values from the data points and from the interval between these points and adds these values to the data set, means that the NPI-B sample has more variance than other methods of bootstrap. We will show this property in detail in simulation studies in this chapter and in the next example. All possible orderings of the new observations among the past observations are equally likely to appear in NPI-B, while they have multinomial distributions with $n+1$ intervals for Banks' bootstrap, and with n data observations for the standard bootstrap, all are equally likely for each new observation.

The NPI-B algorithm for one-dimensional real-valued data on a finite (bounded) interval is as follows:

1. Take the data set of n observations which are real-valued, 1-dimensional on a finite closed interval.
2. These n observations partition the intervals into $n+1$ intervals.
3. Randomly select one of the $n+1$ intervals, each with equal probability.
4. Sample one future value uniformly from this selected interval.
5. Add that value to the data: increase n to $n+1$.
6. Repeat steps 2-4, now with $n+1$ data, to get a further future value.
7. Do this m times to get a NPI bootstrap sample $Y_1, Y_2, ..., Y_m$ of size m.

NPI-B		
Orderings	Frequency	Observed Proportions
(3,0,0,0)	10	0.05
(2,1,0,0)	9	0.05
(2,0,1,0)	11	0.06
(2,0,0,1)	7	0.04
(1,2,0,0)	14	0.07
(1,1,1,0)	11	0.06
(1,1,0,1)	7	0.04
(1,0,2,0)	16	0.08
(1,0,1,1)	10	0.05
(1,0,0,2)	12	0.06
(0,3,0,0)	8	0.04
(0,2,1,0)	9	0.05
(0,2,0,1)	8	0.04
(0,1,2,0)	12	0.06
(0,1,1,1)	7	0.04
(0,1,0,2)	8	0.04
(0,0,3,0)	11	0.06
(0,0,2,1)	11	0.06
(0,0,1,2)	8	0.04
(0,0,0,3)	11	0.06

Table 2.1: Orderings of NPI-B

8. Repeat all these steps B times, where B is a chosen integer value, to get a total of B NPI bootstrap samples of size m.

In this algorithm we assumed that the distribution between data points is uniform, this does not follow from Hill's assumption, but we assumed that because the NPI-B method is an improvement on Banks' bootstrap method. Banks' put uniformly distributed probabilities $1/(n+1)$ over each interval between data points. This assumption is convenient for computation and intuitively reasonable. We do not consider further underlying principles according to which such an assumption would be optimal, it is just one possible assumption among many possibilities.

Example 2.1

In this example we illustrate the main arguments of the three bootstrap methods and the differences between them. We use $(2, 4, 6)$ as original sample, and treat it as a sample drawn from an unknown distribution with support [0,8]. First, to sample an NPI-B sample of size $m = n = 3$ there is $n + 1$ intervals between the data set values including the end points $(0, 8)$. The intervals are $I_1 = (0, 2)$, $I_2 = (2, 4)$, $I_3 = (4, 6)$ and $I_4 = (6, 8)$. Choose one interval and then sample the new value from this interval as the first value in NPI-B sample. Then add this value to the

Banks-B			
Orderings	Theoretical Probabilities	Frequency	Observed Proportions
(3,0,0,0)	0.02	7	0.04
(2,1,0,0)	0.05	7	0.04
(2,0,1,0)	0.05	14	0.07
(2,0,0,1)	0.05	12	0.06
(1,2,0,0)	0.05	9	0.05
(1,1,1,0)	0.09	20	0.10
(1,1,0,1)	0.09	14	0.07
(1,0,2,0)	0.05	5	0.03
(1,0,1,1)	0.09	14	0.07
(1,0,0,2)	0.05	7	0.04
(0,3,0,0)	0.02	5	0.03
(0,2,1,0)	0.05	13	0.07
(0,2,0,1)	0.05	12	0.06
(0,1,2,0)	0.05	9	0.05
(0,1,1,1)	0.09	18	0.09
(0,1,0,2)	0.05	8	0.04
(0,0,3,0)	0.02	2	0.01
(0,0,2,1)	0.05	12	0.06
(0,0,1,2)	0.05	10	0.05
(0,0,0,3)	0.02	2	0.01

Table 2.2: Orderings of Banks-B

data set so that it is $n = 4$ and the intervals become 5 intervals. Continue with this procedure to derive an NPI-B sample of size $m = 3$. There are $\binom{n+m}{m} = \binom{6}{3} = 20$ orderings of 3 future observations among the 3 data observations, which are shown in Table 2.1. For example, $(1, 0, 2, 0)$ means there is 1 future observation from I_1, 0 from I_2, 2 from I_3 and 0 from I_4. All orderings have equal probability $1/20 = 0.05$. We sampled 200 NPI-B samples to record the number of frequencies of each ordering and put the results of the simulation in Table 2.1. It is clear from this table that the probability of each ordering is close to 0.05 in most cases.

In Banks' bootstrap we use the same method but do not add the new value to the data set. Table 2.2 shows the orderings and probability of each one using multinomial distribution, and the observed proportions of each ordering using simulation with 200 Banks' bootstrap samples. For a standard bootstrap sample, the value is drawn just from the data values. This sample can be, for example, $(2, 4, 6)$ or $(2, 2, 6)$ or $(4, 2, 4)$ etc. There are 10 orderings that can appear here, as shown in Table 2.3. This table contains the probability of the ordering using a multinomial distribution with n data observations, and the actual probabilities of 200 standard bootstrap samples. The theoretical probabilities and those from the simulation study are similar in most cases of the three kinds of bootstrap methods. Figure 2.1 illustrates the variance

standard-B			
Orderings	Theoretical Probabilities	Frequency	Observed Proportions
(3,0,0)	0.04	10	0.05
(2,1,0)	0.11	19	0.10
(2,0,1)	0.11	17	0.09
(1,2,0)	0.11	24	0.12
(1,1,1)	0.22	45	0.23
(1,0,2)	0.11	30	0.15
(0,3,0)	0.04	2	0.01
(0,2,1)	0.11	27	0.14
(0,1,2)	0.11	20	0.10
(0,0,3)	0.04	6	0.03

Table 2.3: Orderings of Standard-B

values of NPI-B samples, standard-B samples and Banks' bootstrap samples to measure how far observations are spread out, and to give a general insight into the NPI-B samples that have a large variance. That is due to the method of sampling as discussed earlier. These values of variances come from the simulation experiment in this example and are plotted in Figure 2.1. There are some NPI-B samples which have small values of variance,and some of them are close to 0, as shown in Figure 2.1. This is possible with NPI-B samples but happens rarely. This can appear because the sample size is small.

2.3 NPI Bootstrap for Finite Intervals

The proposed nonparametric predictive inference bootstrap method (NPI-B) is based on the repeated application of assumption $A_{(n)}$. First it is done with the n observed data which create $n+1$ intervals, leading to one further observation. This is followed by adding a further observation to the data and applying $A_{(n+1)}$ in order to draw the second further observation, and so on. This is continued until there are m further observations, which then together (and without the original n observed data) form one NPI-B sample. In this section we restrict attention to NPI-B applied to observations on a finite interval, because it simplifies the approach in the intervals I_1 and I_{n+1}. If NPI-B is applied on the full real-line, these two intervals require different procedures for sampling a value within them. This will be presented in the following sections. The NPI-B algorithm for one-dimensional real-valued data on a finite interval is shown in Section 2.2.

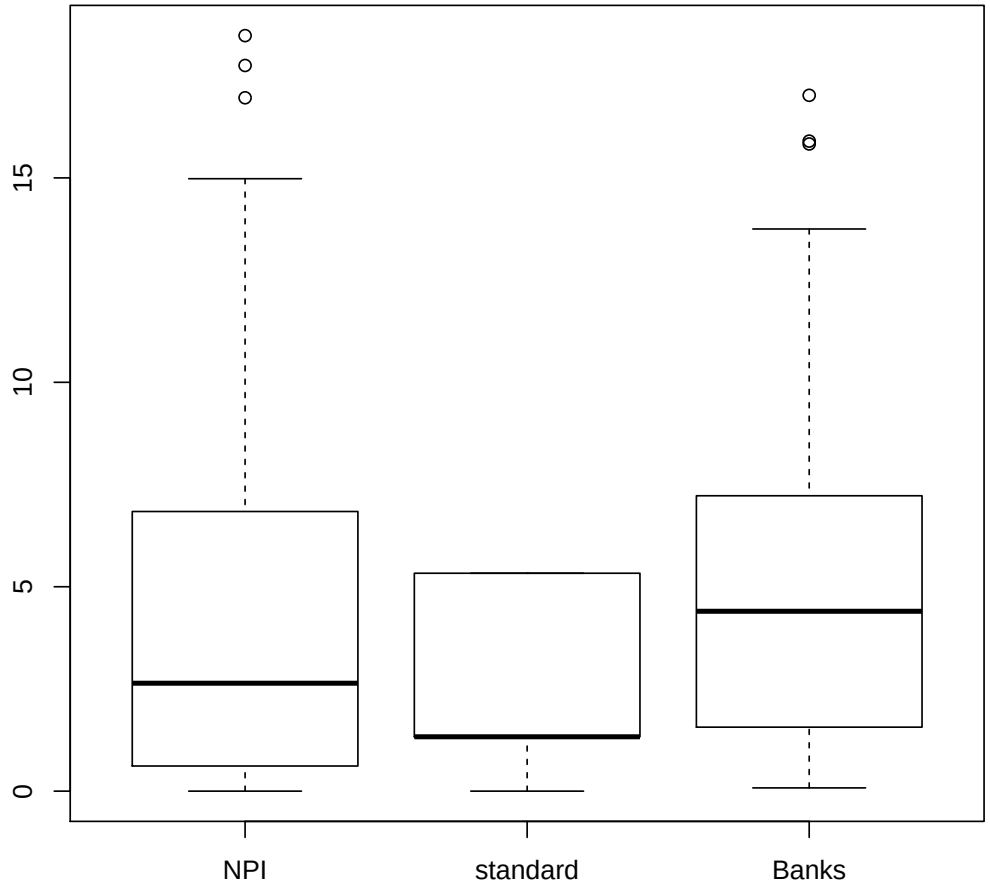

Figure 2.1: Variance of three types of bootstrap methods

The crucial difference from the standard bootstrap ('standard-B') is that an NPI-B sample does not consist of the observations from the original sample but of points from the whole possible data range, because the sampling here is from the interval in between the data values and also outside the data range. This procedure leads to greater variation in the NPI-B samples than in the standard-B samples, as discussed in Section 2.2. Furthermore with NPI-B it is possible to estimate $P(Y > c)$ for different values of c, especially if c is greater than the maximum value of the original sample. This provides a method for estimating this probability, but the uniformity assumption within the intervals will affect the estimate and hence it would be difficult to arrive at the exact statistical properties for such an estimate, hence the results presented mainly serve as an illustration of our method. It should be mentioned that the NPI-B procedure is close in nature to Banks' proposal of a smoothed bootstrap [8], where sampling also takes place uniformly in intervals between the actual data, but Banks' only uses the actual $n + 1$ intervals from step 1 above for the sampling of all values in the bootstrap sample, so a sampled value is not added to the data. This leads to a smaller variation in Banks' approach than in NPI-B.

To study the NPI bootstrap performance, we have carried out a simulation experiment using R code. The considered methods in this part are: standard bootstrap, Banks' bootstrap (smoothed Efron's bootstrap) and NPI bootstrap. We have performed simulation studies, using the software R, for the performance of the NPI-B as the estimation approach. For each method, we generated $B = 1000$ bootstrap samples and calculated the variance which is the square of equation (1.6), bias from equation (1.9), absolute error $|T_n^* - T_n^o|$. T_n^o here is the observed value of statistic which is calculated from the original sample. We found the absolute error for every value of statistics in bootstrap samples and then took the average of these values, and the mean square error (MSE) of the statistics from equation (1.12). It is important to mention the reason for choosing these measures. The variance of statistics is used to show the difference between the three methods of bootstrap or to show which method has a close variance to the original observations. We will see that the NPI-B has the largest variance of statistics but it is the closest one to the

n	Uniform (0,1)	Beta (1,2)	Beta (0.5,1)
20	0.0800	0.0551	0.0729
50	0.0815	0.0444	0.0738
100	0.0857	0.0523	0.0799
200	0.0853	0.0492	0.0825
500	0.0842	0.0523	0.0859
1000	0.0841	0.0521	0.0834

Table 2.4: Variances of original samples from specific distributions with a variety of sample sizes

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of mean	0.004	0.002	0.001	0.0005	0.0002	0.0001
	bias	0.003	-0.0001	-0.001	-0.001	0.0002	0.0001
	absolute error	0.049	0.032	0.024	0.017	0.010	0.007
	MSE	0.004	0.002	0.001	0.0005	0.0001	0.0001
Banks	variance of mean	0.004	0.002	0.001	0.0004	0.0001	0.001
	bias	-0.007	0.001	0.0003	0.001	-0.0001	-0.0001
	absolute error	0.052	0.032	0.023	0.017	0.010	0.007
	MSE	0.004	0.002	0.001	0.0004	0.0002	0.0001
NPI	variance of mean	0.008	0.003	0.002	0.001	0.0003	0.0002
	bias	-0.009	0.002	0.001	0.001	-0.0002	0.00001
	absolute error	0.070	0.046	0.033	0.024	0.015	0.010
	MSE	0.008	0.003	0.002	0.001	0.0003	0.0002

Table 2.5: The sample mean when the original sample was from U (0,1)

variance of the original sample. Regarding bias, MSE and absolute error, they are the most commonly used measures of statistical accuracy of estimators as discussed in Section 1.3.1. We repeated this experiment for $n = 20, 50, 100, 200, 500, 1000$ and for statistics: mean, variance and upper quartile (q_{75}), to present location, variation and position parameters, respectively, and for different distributions such as Uniform (0,1), Beta (α, β), with $(\alpha = 0.5, \beta = 1)$ and $(\alpha = 1, \beta = 2)$, as examples of symmetric and skewed distributions. Note that, in these cases, the size of bootstrap samples is $m = n$. Table 2.4 shows the observed variance of these original samples with various sample sizes.

Tables 2.5, 2.6, 2.7, of Uniform (0,1) show that when comparing the absolute value of bias, the NPI bootstrap has the smallest bias of variance parameter in all cases, but for the mean and q_{75} it has the smallest value only if n is very large (500 or 1000). Mostly, for three parameters, the MSE and absolute error of Banks' bootstrap and standard bootstrap are very close but they are less than NPI bootstrap's values of MSE and absolute error. The variance of three parameters using NPI-B is larger than the others methods. This property of NPI-B is considered a good point because

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of variance	0.0003	0.0001	0.0001	0.00003	0.00001	0.00001
	bias	-0.005	-0.002	-0.001	-0.0005	-0.0001	-0.0001
	absolute error	0.015	0.009	0.006	0.004	0.003	0.002
	MSE	0.0003	0.0001	0.0001	0.00003	0.00001	0.00001
Banks	variance of variance	0.0004	0.0001	0.0001	0.00003	0.00001	0.00001
	bias	0.005	0.002	0.001	0.0004	0.0002	0.00004
	absolute error	0.016	0.009	0.007	0.004	0.003	0.002
	MSE	0.0004	0.0001	0.0001	0.00003	0.00001	0.00001
NPI	variance of variance	0.001	0.0002	0.0001	0.0001	0.00003	0.00001
	bias	0.001	0.0002	-0.0001	-0.00002	0.00002	-0.0001
	absolute error	0.021	0.012	0.009	0.006	0.004	0.003
	MSE	0.001	0.0002	0.0001	0.0001	0.00003	0.00001

Table 2.6: The sample variance when the original sample was from U (0,1)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of q_{75}	0.004	0.003	0.001	0.001	0.0003	0.0002
	bias	-0.006	-0.014	-0.006	-0.004	0.0005	0.001
	absolute error	0.039	0.040	0.020	0.024	0.014	0.009
	MSE	0.004	0.003	0.001	0.001	0.0003	0.0002
Banks	variance of q_{75}	0.004	0.002	0.001	0.001	0.0003	0.0002
	bias	0.001	-0.007	-0.003	-0.000	0.001	0.001
	absolute error	0.044	0.037	0.018	0.023	0.015	0.009
	MSE	0.004	0.002	0.001	0.001	0.0003	0.0002
NPI	variance of q_{75}	0.008	0.005	0.001	0.001	0.001	0.0004
	bias	-0.009	-0.012	-0.004	-0.002	0.0002	0.001
	absolute error	0.064	0.053	0.028	0.030	0.020	0.014
	MSE	0.008	0.005	0.001	0.001	0.001	0.0004

Table 2.7: The sample upper quartile q_{75} when the original sample was from U (0,1)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of mean	0.003	0.001	0.001	0.0003	0.0001	0.0001
	bias	0.001	-0.0005	-0.002	0.0002	0.0002	0.0003
	absolute error	0.041	0.024	0.019	0.013	0.008	0.006
	MSE	0.003	0.001	0.001	0.0003	0.0001	0.0001
Banks	variance of mean	0.006	0.002	0.001	0.0003	0.0001	0.0001
	bias	0.078	0.035	0.018	0.009	0.003	0.002
	absolute error	0.089	0.041	0.024	0.015	0.009	0.006
	MSE	0.012	0.003	0.001	0.0004	0.0001	0.0001
NPI	variance of mean	0.011	0.003	0.001	0.001	0.0002	0.0001
	bias	0.076	0.036	0.018	0.009	0.003	0.002
	absolute error	0.099	0.050	0.031	0.020	0.012	0.008
	MSE	0.017	0.004	0.002	0.001	0.0002	0.0001

Table 2.8: The sample mean when the original sample was from Beta (1,2)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of variance	0.0002	0.0001	0.0001	0.00002	0.00001	0.000004
	bias	-0.003	-0.001	-0.001	-0.0003	-0.0002	0.0001
	absolute error	0.013	0.008	0.005	0.004	0.002	0.002
	MSE	0.0003	0.0001	0.0001	0.00002	0.00001	0.000004
Banks	variance of variance	0.005	0.001	0.0003	0.0001	0.00002	0.00001
	bias	0.068	0.033	0.016	0.008	0.003	0.001
	absolute error	0.073	0.035	0.017	0.009	0.004	0.002
	MSE	0.010	0.002	0.001	0.0002	0.00003	0.00001
NPI	variance of variance	0.009	0.002	0.001	0.0002	0.00004	0.00001
	bias	0.062	0.032	0.015	0.008	0.003	0.001
	absolute error	0.072	0.036	0.018	0.010	0.005	0.003
	MSE	0.013	0.003	0.001	0.0002	0.0001	0.00002

Table 2.9: The sample variance when the original sample was from Beta (1,2)

it makes the variance of parameters using NPI-B the closest one to the variance of the original sample, and supports the discussion in Section 2.2. In the tables of results in this chapter the values were approximated to three decimal digits to make them easy to read, but with some values we use additional digits to be informative and to avoid putting zero's "0.000" in the results. It is apparent from results of Beta (1,2) in Tables 2.8, 2.9, 2.10 that for all parameters here the bias of standard-B is the smallest one, unlike the other two methods, but sometimes there is no large difference between the bias of NPI-B and the bias of Banks' bootstrap method.

Here, for the mean and q_{75}, note that the absolute error and MSE have the same status of Uniform (0,1), but for variance, the absolute error of Banks' bootstrap and NPI bootstrap are similar and greater than the standard bootstrap.

The results of Beta (0.5,1) are presented in Tables 2.11,2.12,2.13 and show that when we estimate the variance, the NPI bootstrap method has the smallest bias in most cases, but for the mean and q_{75} it does not perform better than other

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of q_{75}	0.007	0.003	0.001	0.001	0.0004	0.0002
	bias	0.013	-0.010	-0.004	0.002	-0.0003	0.002
	absolute error	0.067	0.047	0.029	0.021	0.017	0.012
	MSE	0.007	0.003	0.001	0.001	0.0004	0.0002
Banks	variance of q_{75}	0.016	0.003	0.001	0.001	0.0005	0.0002
	bias	0.082	0.015	0.010	0.009	0.002	0.002
	absolute error	0.109	0.048	0.029	0.023	0.018	0.012
	MSE	0.022	0.003	0.001	0.001	0.0005	0.0002
NPI	variance of q_{75}	0.028	0.007	0.003	0.002	0.001	0.0004
	bias	0.088	0.015	0.009	0.009	0.001	0.002
	absolute error	0.132	0.065	0.040	0.032	0.024	0.016
	MSE	0.036	0.007	0.003	0.002	0.001	0.0004

Table 2.10: The sample upper quartile q_{75} when the original sample was from Beta (1,2)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of mean	0.003	0.001	0.001	0.0004	0.0002	0.0001
	bias	-0.003	-0.001	0.0005	-0.00003	-0.0002	0.0001
	absolute error	0.047	0.029	0.022	0.002	0.011	0.007
	MSE	0.003	0.001	0.001	0.0004	0.0002	0.0001
Banks	variance of mean	0.004	0.002	0.001	0.0004	0.0002	0.0001
	bias	0.030	0.012	0.006	0.003	0.001	0.001
	absolute error	0.054	0.033	0.023	0.017	0.010	0.007
	MSE	0.005	0.002	0.001	0.0004	0.0002	0.0001
NPI	variance of mean	0.007	0.003	0.002	0.001	0.0003	0.0002
	bias	0.030	0.011	0.006	0.003	0.001	0.001
	absolute error	0.071	0.044	0.032	0.023	0.015	0.010
	MSE	0.008	0.003	0.002	0.001	0.0003	0.0002

Table 2.11: The sample mean when the original sample was from Beta (0.5,1)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of variance	0.0004	0.0002	0.0001	0.0001	0.00002	0.00001
	bias	-0.004	-0.002	-0.001	-0.001	-0.0002	-0.0001
	absolute error	0.016	0.011	0.009	0.006	0.004	0.002
	MSE	0.0004	0.0002	0.0001	0.0001	0.00002	0.00001
Banks	variance of variance	0.0004	0.0002	0.0001	0.00004	0.00002	0.00001
	bias	-0.002	0.001	0.001	0.001	0.0002	0.00003
	absolute error	0.016	0.011	0.009	0.005	0.003	0.002
	MSE	0.0004	0.0002	0.0001	0.00004	0.00002	0.00001
NPI	variance of variance	0.001	0.0004	0.0002	0.0001	0.00004	0.00002
	bias	-0.005	-0.0003	-0.0002	0.0002	0.00002	-0.0001
	absolute error	0.021	0.015	0.012	0.007	0.005	0.003
	MSE	0.001	0.0004	0.0002	0.0001	0.00004	0.00002

Table 2.12: The sample variance when the original sample was from Beta (0.5,1)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance of q_{75}	0.006	0.006	0.002	0.003	0.001	0.0004
	bias	-0.038	0.018	0.007	0.012	0.011	0.001
	absolute error	0.064	0.056	0.029	0.040	0.028	0.016
	MSE	0.007	0.006	0.002	0.003	0.001	0.0004
Banks	variance of q_{75}	0.006	0.005	0.003	0.002	0.001	0.0004
	bias	-0.017	0.037	0.017	0.018	0.013	0.003
	absolute error	0.059	0.060	0.033	0.038	0.027	0.016
	MSE	0.006	0.007	0.003	0.003	0.001	0.0004
NPI	variance of q_{75}	0.015	0.011	0.007	0.004	0.002	0.001
	bias	-0.021	0.030	0.021	0.021	0.015	0.003
	absolute error	0.088	0.082	0.055	0.052	0.036	0.022
	MSE	0.015	0.012	0.007	0.005	0.002	0.001

Table 2.13: The sample upper quartile q_{75} when the original sample was from Beta (0.5,1)

methods in bias. Sometimes the bias of the NPI bootstrap and the bias of the Banks' bootstrap are close, for the mean parameter, whereas for variance the bias of the standard bootstrap and Banks' bootstrap are similar. As we mentioned before, the values of MSE and absolute error are close and less than these measures in the NPI bootstrap. In this section, we studied the NPI bootstrap performance with data from distributions, which have restricted intervals, such as Uniform (0,1), Beta (0.5,1) and Beta (1,2) and the simulation results showed that NPI-B gives the smallest bias when estimating the variance parameter, just with Uniform (0,1), unlike other distributions studied here. Furthermore, the variance of statistics using NPI-B is the greatest and closest to the variance of the original sample which is described in Table 2.4. This is considered a good point of NPI-B because it has a variance that is close to the variance of the underlying distribution. For example, with Uniform(0,1) the variance of mean using NPI-B and $n = 20$ is 0.008, is the largest one and the closest to the variance of the original sample=0.08, and that appears to all distributions in this section.

2.4 NPI Bootstrap on Real Line

Now we want to generalize our method to observations on the real line $(-\infty, +\infty)$ or $[0, \infty)$. These cases require some assumptions. Our method depends on dividing the observations into intervals, and then sampling the observations uniformly from

these intervals. If we have real line data we can not do that with all the intervals, and we can not sample uniformly from $(-\infty, x_1)$ and $(x_n, +\infty)$. Instead we sample observations from $(-\infty, x_1)$ and $(x_n, +\infty)$ by assuming the tails of a normal distribution are these intervals. To estimate the parameters of Normal distribution μ and σ, we take

$$\mu = \frac{x_1 + x_n}{2} \tag{2.1}$$

We know that $P(Y > x_n) = \frac{1}{n+1}$, and by using the properties of normal cumulative function with this probability we can estimate σ

$$P(Y > x_n) = 1 - \Phi\left(\frac{x_n - \mu}{\sigma}\right) = \frac{1}{n+1} \tag{2.2}$$

$$\sigma = \frac{x_n - \mu}{\Phi^{-1}(\frac{n}{n+1})} \tag{2.3}$$

To draw NPI-B samples here we follow the algorithm in Section 2.3, but use a different act with intervals $(-\infty, x_1)$ and $(x_n, +\infty)$. If the chosen interval is $(-\infty, x_1)$ or $(x_n, +\infty)$, we draw the future value from Normal distribution with parameters (μ, σ^2), which are defined in equations (2.1) and (2.3), and we accept this future value if it is larger than x_n (for $(x_n, +\infty)$), and if it is smaller than x_1 (for $(-\infty, x_1)$). Here we are sampling from the conditional tail distribution. It is of course possible to assume a different distribution for the tails, or to use a different method to fit the Normal distribution, but for most inferences this is unlikely to make much difference while the implementation should also be straightforward.

In this section the simulation study considered NPI-B and standard-B without Banks' bootstrap because we could not make assumptions for the end intervals for it. We compare the standard bootstrap method and NPI bootstrap method using the Student's t-distribution and the Normal distribution. For each one we generated $B = 1000$ bootstrap samples from the original sample of different sizes $n = 20, 50, 100, 200, 500, 1000$ from a Student's t-distribution with degrees of freedom 4. To compare, we calculated the bias, MSE and absolute error as we did before. Where T_n^o is not the observed value of a statistic which is calculated from the original sample, here it is the value of parameters from the underlying population. These

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.096	0.035	0.018	0.009	0.004	0.002
	bias	0.641	0.218	0.137	0.157	0.086	0.009
	absolute error	0.643	0.240	0.156	0.160	0.091	0.037
	MSE	0.507	0.083	0.037	0.033	0.011	0.002
NPI	variance	0.272	0.085	0.038	0.019	0.008	0.004
	bias	0.661	0.244	0.145	0.157	0.080	0.007
	absolute error	0.704	0.302	0.195	0.174	0.099	0.053
	MSE	0.709	0.144	0.059	0.044	0.015	0.004

Table 2.14: The sample mean when the original sample was from Student's t_4

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.773	0.221	0.086	0.049	0.062	0.040
	bias	-0.161	-0.378	-0.279	-0.315	-0.020	0.071
	absolute error	0.741	0.505	0.337	0.333	0.197	0.168
	MSE	0.799	0.364	0.164	0.148	0.062	0.045
NPI	variance	4.427	1.217	0.389	0.160	0.192	0.109
	bias	0.842	0.183	-0.011	-0.160	0.134	0.175
	absolute error	1.529	0.799	0.464	0.353	0.330	0.276
	MSE	5.136	1.250	0.389	0.186	0.210	0.140

Table 2.15: The sample variance when the original sample was from Student's t_4

values have a specific formula depending on the distribution of the population. We will start with Student's t-distribution with degrees of freedom $\upsilon = 4$ with mean $\mu = 0$ and variance $\sigma^2 = \upsilon/(\upsilon - 2)$ (if $\upsilon > 2$). Note that, in these cases, the size of bootstrap samples is $m = n$, and the considered statistics here are mean and variance.

From Table 2.14, we can see that the absolute value of bias in the NPI bootstrap samples is larger than the same measure in standard bootstrap samples except for $n = 500, 1000$. The MSE and absolute error for the mean in NPI bootstrap method are the largest. This status of MSE and absolute error accrues also for variance in Table 2.15, and the bias in our method is the largest only in cases $n = 50, 100, 200$. By comparing the values of variance of statistics in two methods, the NPI bootstrap method has the largest variance and is closer to the variance of underlying distribution Students'$t(4) = 2$. This is a positive point of the NPI-B method and occurs when we study it in restricted intervals such as Uniform distribution.

Note that in our NPI bootstrap samples there is a possibility for the occurrence of some values larger than x_n in the original sample, because we choose our sample from the observations and from the intervals between observations $(-\infty, x_1), (x_1, x_2), ...,$ $(x_n, +\infty)$. But in the standard bootstrap we sample values only from the observa-

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.153	0.060	0.029	0.016	0.007	0.003
	bias	-0.659	-0.281	-0.058	-0.022	-0.045	-0.053
	absolute error	0.678	0.315	0.144	0.102	0.074	0.065
	MSE	0.587	0.139	0.033	0.016	0.009	0.006
NPI	variance	0.402	0.141	0.065	0.036	0.014	0.007
	bias	-0.626	-0.266	-0.059	-0.016	-0.045	-0.053
	absolute error	0.743	0.373	0.209	0.152	0.103	0.081
	MSE	0.794	0.212	0.069	0.036	0.016	0.010

Table 2.16: The sample mean when the original sample was from Normal (28,4)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.878	0.402	0.174	0.095	0.050	0.027
	bias	-1.115	-1.048	-1.031	-0.544	-0.387	-0.318
	absolute error	1.252	1.083	1.036	0.554	0.395	0.322
	MSE	2.122	1.499	1.238	0.391	0.199	0.128
NPI	variance	4.732	1.458	0.513	0.228	0.118	0.062
	bias	0.004	-0.527	-0.801	-0.408	-0.299	-0.261
	absolute error	1.627	1.079	0.925	0.523	0.382	0.304
	MSE	4.732	1.735	1.155	0.394	0.207	0.130

Table 2.17: The sample variance when the original sample was from Normal (28,4)

tions. The NPI bootstrap method for all parameters studied here has the largest variance which is closer to the variance of underlying distribution.
Now we repeat the experiment with the same assumptions but with the original sample from Normal distribution with parameters mean $\mu = 28$ and variance $\sigma^2 = 4$, and the statistics which we want to study are: mean and variance.

Tables 2.16 and 2.17 show that, when we estimate the mean, the NPI bootstrap method has the smallest bias in most cases. For the variance it works very well for all cases in bias, and the absolute error has the smallest value for all cases, except in $n = 20$. But for the mean it does not perform better than the standard bootstrap method. The values of MSE in the standard bootstrap method are the smallest. Note that, for all cases and all events which we studied in Normal $(28, 4)$, the variance of statistics in our method is greater than the variance in the standard bootstrap and closer to the variance of normal distribution which is 4, as we mentioned before, when we discussed NPI bootstrap with Students't-distribution.

In the case of the data on $[0, \infty)$, for example lifetime data, we need to define a specific distribution for sampling from $(x_n, +\infty)$. We use the tail of an Exponential

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.031	0.009	0.004	0.002	0.001	0.0004
	bias	0.052	0.044	0.013	-0.066	-0.050	-0.029
	absolute error	0.146	0.085	0.053	0.069	0.051	0.030
	MSE	0.034	0.011	0.004	0.006	0.003	0.001
NPI	variance	0.143	0.030	0.011	0.004	0.002	0.001
	bias	0.164	0.086	0.031	-0.057	-0.048	-0.027
	absolute error	0.288	0.142	0.087	0.072	0.053	0.032
	MSE	0.170	0.037	0.012	0.007	0.004	0.001

Table 2.18: The sample mean when the original sample was from Weibull (1.5,1)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.114	0.023	0.008	0.003	0.001	0.0004
	bias	0.266	0.139	0.100	0.025	-0.002	-0.018
	absolute error	0.334	0.163	0.110	0.049	0.025	0.023
	MSE	0.185	0.043	0.018	0.004	0.001	0.001
NPI	variance	3.442	0.497	0.053	0.020	0.004	0.001
	bias	0.908	0.354	0.186	0.070	0.015	-0.010
	absolute error	0.986	0.393	0.202	0.101	0.045	0.029
	MSE	4.267	0.623	0.088	0.025	0.004	0.001

Table 2.19: The sample variance when the original sample was from Weibull (1.5,1)

distribution for this. In this case the intervals will be $(0, x_1), (x_1, x_2), ..., (x_n, +\infty)$. If the chosen interval is $(x_n, +\infty)$ we select the observation randomly from Exponential distribution. Otherwise we draw the observation uniformly from other intervals. To estimate the parameters of Exponential distribution λ, we use the cumulative function $P(Y < y) = 1 - e^{-\lambda y}$. We know that $P(Y > x_n) = \frac{1}{n+1}$, so we get

$$\lambda = \frac{\ln(n+1)}{x_n} \tag{2.4}$$

A different distribution, or a different way to fit the Exponential distribution, could be assumed for this tail, we have not investigated this further but expect that it would not make much difference for most inferences. To sample NPI-B samples in this case we use the algorithm in Section 2.2, but if the chosen interval is $(x_n, +\infty)$, we draw the future value from Exponential distribution with parameters λ in equation (2.4) and we accept this future value if it is larger than x_n, we sample here from the conditional tail distribution.

To study this case we started with the original sample from Weibull distribution with parameters shape α and scale β, $\alpha = 1.5, \beta = 1$ and considered bootstrap for the mean and variance.

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.109	0.210	0.105	0.048	0.016	0.007
	bias	-1.066	-0.303	0.068	0.183	0.100	0.055
	absolute error	1.066	0.454	0.261	0.231	0.133	0.082
	MSE	1.245	0.302	0.110	0.081	0.026	0.010
NPI	variance	0.379	0.695	0.262	0.106	0.034	0.016
	bias	-0.962	-0.077	0.171	0.237	0.122	0.067
	absolute error	1.019	0.649	0.409	0.323	0.175	0.115
	MSE	1.304	0.701	0.291	0.163	0.048	0.021

Table 2.20: The sample mean when the original sample was from Gamma (2,2)

method	measures	$n = 20$	$n = 50$	$n = 100$	$n = 200$	$n = 500$	$n = 1000$
standard	variance	0.377	30.167	9.311	2.836	0.663	0.275
	bias	-5.724	2.571	2.321	1.657	0.358	-0.105
	absolute error	5.724	4.738	2.983	1.911	0.704	0.429
	MSE	33.141	36.778	14.699	5.583	0.791	0.286
NPI	variance	15.594	373.136	70.245	19.307	2.560	0.815
	bias	-4.411	9.791	5.518	3.094	0.911	0.176
	absolute error	5.183	11.902	6.334	3.504	1.328	0.701
	MSE	35.055	468.998	100.696	28.881	3.390	0.845

Table 2.21: The sample variance when the original sample was from Gamma (2,2)

It is clear from Tables 2.18 and 2.19 that the bias of the NPI bootstrap method is not the smallest unless when $n \geq 200$ for the mean and when $n = 1000$ for the variance. For mean and variance, note that the absolute error and MSE of the NPI bootstrap method are greater than the same measures of the standard bootstrap method. Note that the variance of statistics using NPI-B has the largest value as the case of Normal and Students't-distributions.

Tables 2.20 and 2.21 show the experiment with the original sample from the Gamma distribution with parameters shape α and scale β, $\alpha = 2, \beta = 2$ and statistics: mean and variance.

The bias of the NPI bootstrap method, for the mean, is the smallest when $n = 20, 50$ and for the variance when $n = 20$. Otherwise the values of the bias in our method are larger than the bias of the standard bootstrap. The absolute error of the NPI bootstrap method, for mean and variance, is very large except when $n = 20$ and then it has small values. The variance of all the statistics in the NPI bootstrap method is the largest and the closest to the variance of Gamma $(2, 2)$ which is 8, but for the variance parameter the variance is sometimes overestimated.

	Student't(4)					
$m = n =$	20	50	100	200	500	1000
no. of min	477	500	487	496	518	529
no. of max	496	530	513	482	470	490
no. of between	18183	48000	98025	198000	498043	997996
	Normal(28,4)					
$m = n =$	20	50	100	200	500	1000
no. of min	492	484	478	495	504	487
no. of max	526	486	489	511	518	533
no. of between	18007	48053	98075	197966	497924	997919
	Weibull(1.5,1)					
$m = n =$	20	50	100	200	500	1000
no. of min	491	517	488	512	483	497
no. of max	496	488	528	497	517	500
no. of between	18057	48060	98008	197983	497911	997951
	Gamma(2,2)					
$m = n =$	20	50	100	200	500	1000
no. of min	496	530	525	483	507	466
no. of max	487	492	504	507	507	533
no. of between	18096	48049	97901	198026	497914	998009

Table 2.22: Number of observations within data set

In general the NPI-B samples have more variations than those in the standard and Banks' bootstrap samples. On the other hand they have less bias than the other two bootstrap methods, in some cases, when the variance parameter is estimated.

Table 2.22 shows how many NPI bootstrap samples have the future minimum value Y_{n+1} smaller than a minimum value of the original sample x_1 (no. of min), and how many samples have the future maximum value Y_{n+m} larger than original maximum value x_n (no. of max). The third row contains the number of observations of NBI-B samples between minimum and maximum values x_1, x_n, for Normal, Students't, Gamma and Weibull distribution. For example, in Students't-distribution with degrees of freedom 4 when $n = 20$ and $B = 1000$ there are 477 NPI-B samples from 1000 that have a minimum value smaller than the original minimum value.

As we mentioned before, NPI is a statistical method based on Hill's assumption A_n:

$$P(Y_j < Y_{n+1} < Y_{j+1}) = \frac{1}{n+1}, \quad j = 0, 1, ..., n \tag{2.5}$$

This means that the probability that a future value lies in the interval (Y_j, Y_{j+1}) is $1/(n+1)$. So the probability that the NPI bootstrap sample has all values

smaller than the original maximum is 1/2. If we have m future values $Y_{n+1}, ..., Y_{n+m}$ then the probability that these m values are smaller than the original maximum is:$\frac{n}{n+1}.\frac{n+1}{n+2}.\frac{n+2}{n+3}...\frac{n+m-1}{n+m}$, that is equal to $P(Y_{n+1} < X_n) \times P(Y_{n+2} < X_n | Y_{n+1} < X_n) \times P(Y_{n+3} < X_n | Y_{n+1}, Y_{n+2} < X_n) \times ... \times P(Y_{n+m} < X_n | Y_{n+1}, Y_{n+2}...Y_{n+m-1} < X_n)$. And if $m = n$, then the probability will be :$\frac{n}{n+1}.\frac{n+1}{n+2}.\frac{n+2}{n+3}...\frac{2n-1}{2n} = \frac{n}{2n} = \frac{1}{2}$. We can see this result in Table 2.22. For example, for samples from Student't-distribution when $n = 20$ the probability is equal to $\frac{477}{1000} = 0.477$. It is very close to 0.5 and in some cases it is equal to 0.5. Similarly, we know that

$$P(Y_j < Y_{n+1} < Y_{j+k}) = \frac{k}{n+1}, \quad k = 1, ..., n-j+1 \tag{2.6}$$

This is the probability that the future value lies between Y_j and Y_{j+k} which contains k intervals. The probabability for the event that every observation in one NPI bootstrap sample is between the original minimum and original maximum $\frac{k}{n+1}.\frac{k+1}{n+2}.\frac{k+2}{n+3}...\frac{k+m-1}{n+m}$ and $k = n-1$. So we can write the probability as $\frac{n-1}{n+1}.\frac{n}{n+2}.\frac{n+1}{n+3}...\frac{n+m-2}{n+m}$ and if $m = n$ the last probability will be $\frac{n-1}{n+1}.\frac{n}{n+2}.\frac{n+1}{n+3}...\frac{2n-2}{2n} = \frac{n(n-1)}{2n(2n-1)} = \frac{1}{2}\frac{n-1}{2n-1}$ which is close to 0.25. This occurred, for example, for sample from Student't-distribution when $n = 20$ there are 17 observations between the original minimum and maximum of one bootstrap sample from 1000 samples. This is equal to: $\frac{19}{21}.\frac{20}{22}.\frac{21}{23}...\frac{34}{36}.\frac{35}{37} = 0.28$, which is not equal to 0.25 but close to it.

From Table 2.22 we can see that approximately 2000 values were either less than X_1 or greater than X_n, for example when $n = 20$, $20000 - 18183 = 1817$. Let Z_i be a random quantity that is 1 if the $i-th$ NPI-Bootstrap observation is outside $(X_1, ..., X_n)$ and 0 otherwise, Z_i's are independent

$$Z_i = \begin{cases} 1 & \text{with probability } \frac{2}{n+1} \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

the number of observations outside $(X_1, ..., X_n)$ in the jth NPI-bootstrap sample is $w_j = \sum_{i=1}^{n} Z_i$, $E(w) = E(\sum_{i=1}^{n} Z_i) = \sum_{i=1}^{n} E(Z_i) = n.\frac{2}{n+1}$. And the number of observations in B NPI-bootstrap samples and outside $(X_1, ..., X_n)$ is $\sum_{j=1}^{B} w_j$, $E(\sum_{j=1}^{B} w_j) = \sum_{j=1}^{B} E(w_j) = 2.B.\frac{n}{n+1} \approx 2000$, because $B = 1000$ and $\frac{n}{n+1} \approx 1$.

2.5 Comparison With Other Methods

This section considers the comparison between standard-B and NPI-B using confidence intervals and prediction intervals, in order to investigate their performance in estimation and prediction inference.

2.5.1 Confidence Intervals

A comparison between NPI-B and standard-B using bias, MSE and absolute error is not easy because there are different values and different conclusions every time we run the program. So the difficulty of point estimates make us think about using the confidence intervals to compare different kinds of bootstrap methods in estimation matter. We showed some details about bootstrap confidence intervals in Section 1.3.4, and will choose BCa interval in equation (1.23) to compare methods because it has two advantages: a high order of accuracy and transformation respecting, [38]. We could not use the ready code in R, so we wrote a new program to do this task. In this study we use Uniform (0,1), Normal (28,4) and Gamma (2,1) with a different original sample size n and a bootstrap sample size m, $m = n = 20, 50, 100, 200, 500$, for $\alpha = 0.01, 0.05$. To illustrate the performance of $(1 - 2\alpha)$th BCa interval we construct 1000 intervals of each kind and resample $B = 1000$ bootstrap samples each time.

For Uniform (0,1), in Table 2.23, we constructed BCa intervals for mean, variance and q_{75}. We notice that, when $\alpha = 0.05$ and $1 - 2\alpha = 0.90$, the results of NPI-B show some undercoverage for three parameters and for all cases of n. Undercoverage also occurs for the standard-B for the mean when $n = 20$ but it is better than NPI-B in other cases, because it has the presumed coverage probability 0.90 with $n = 20, 50, 100$. Otherwise, the standard-B show some undercoverage results for variance and q_{75} when $n = 20, 50, 200$. When $\alpha = 0.01$ and $1 - 2\alpha = 0.98$, for the mean and q_{75}, the standard-B performs better than NPI-B because it has a higher coverage proportion.

For Gamma (2,1) in Table 2.24, when $\alpha = 0.05, 1 - 2\alpha = 0.90$, the standard-B has more high coverage proportion than NPI-B, for mean, variance and q_{75}, but

	mean									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.96	0.96	0.96	0.95	0.96	0.69	0.71	0.69	0.70	0.73
standard-B	0.98	0.98	0.98	0.97	0.98	0.90	0.90	0.90	0.87	0.91
	variance									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.98	0.96	0.97	0.96	0.73	0.74	0.70	0.68	0.69
standard-B	0.95	0.98	0.98	0.98	0.97	0.88	0.89	0.90	0.89	0.90
	q_{75}									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.96	0.97	0.96	0.95	0.96	0.74	0.72	0.70	0.71	0.74
standard-B	0.97	0.97	0.98	0.98	0.98	0.88	0.89	0.90	0.89	0.91

Table 2.23: Coverage of $(1 - 2\alpha)$ confidence interval of properties of Uniform (0,1)

	mean									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.93	0.95	0.94	0.95	0.96	0.62	0.64	0.66	0.64	0.66
standard-B	0.95	0.97	0.97	0.98	0.98	0.87	0.89	0.89	0.89	0.89
	variance									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.94	0.91	0.92	0.93	0.92	0.62	0.60	0.62	0.58	0.58
standard-B	0.84	0.91	0.95	0.95	0.96	0.76	0.83	0.83	0.86	0.88
	q_{75}									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.96	0.97	0.95	0.96	0.96	0.72	0.72	0.74	0.70	0.72
standard-B	0.97	0.98	0.98	0.98	0.97	0.89	0.90	0.89	0.89	0.89

Table 2.24: Coverage of $(1 - 2\alpha)$ confidence interval of properties of Gamma (2,1)

	mean									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.98	0.97	0.95	0.95	0.69	0.70	0.70	0.68	0.70
standard-B	0.97	0.98	0.98	0.97	0.99	0.87	0.90	0.90	0.89	0.91
	variance									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.92	0.90	0.93	0.92	0.95	0.59	0.55	0.56	0.58	0.62
standard-B	0.91	0.96	0.97	0.98	0.98	0.83	0.87	0.88	0.90	0.91
	q_{75}									
	$\alpha = 0.01$					$\alpha = 0.05$				
$m = n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.97	0.97	0.96	0.96	0.73	0.71	0.71	0.69	0.71
standard-B	0.97	0.98	0.98	0.98	0.98	0.89	0.87	0.90	0.90	0.92

Table 2.25: Coverage of $(1 - 2\alpha)$ confidence interval of properties of Normal (28,4)

they show undercoverage results in most cases. When $\alpha = 0.01$, $1 - 2\alpha = 0.98$, the standard-B for the mean performs better than NPI-B, but sometimes the results of standard-B show undercoverage. Standard-B has the nominal coverage probability 0.98 when $n = 200, 500$ for the mean. For q_{75},the standard-B performs better than NPI-B in all cases, and the nominal coverage probability appears in standard-B when $n = 50, 100, 200$. For variance the NPI-B has the smallest values of coverage proportions in most cases.

Table 2.25 shows the result of Normal (28,4) and $\alpha = 0.05$, $1 - 2\alpha = 0.90$. The NPI-B shows undercoverage results for three parameters and also the standard-B for variance and mean when $n = 20, 200$. For q_{75}, undercoverage occurs in most cases with standard-B except when $n = 100, 200, 500$. When $\alpha - 0.01$, $1 - 2\alpha = 0.98$, the NPI-B shows undercoverage for variance but it works well for the mean and q_{75}, the standard-B is better for the mean and q_{75}.

From the above we see that the NPI-B does not perform well in confidence intervals and that means it is not a good method for estimation. As discussed earlier the NPI-B samples have a larger variance so we expect that the NPI-B intervals will be wider than the standard-B intervals. So why do they have worse coverage? The reason is that the variation of midpoints in NPI-B intervals is greater than in standard-B intervals. For example with Normal (28,4) and $\alpha = 0.05$, $n = 20$, the coverage probability of q_{75} using NPI-B is 0.63, but when using standard-B it is 0.85.

When studying the midpoints of each kind of interval we found that the variance of midpoints of NPI-B intervals is 0.68, while for standard-B intervals it is 0.35. For the variance, the coverage of NPI-B is 0.60 and of standard-B it is 0.78, and the variance of midpoints of NPI-B intervals is 33.26 and of standard-B it is 3.18. But for the mean, the variance of midpoints of intervals of two types of bootstrap is similar. It is 0.26, while the coverage of NPI-B is 0.60 and of standard-B it is 0.83.

2.5.2 Prediction Intervals

The NPI-B method is considered as a prediction approach because it depends on assumption $A_{(n)}$ for prediction, so next we discuss this aspect. There are some illustrations in Section 1.3.4 which show prediction intervals for observations and for parameters. Those intervals show how to test the prediction performance of NPI-B. We start with the percentile interval to predict m future observations and consider the NPI-B sample as the future observations as done in [55], but without bootstrap calibration. We follow the steps:

1. Draw c (say c=100) original samples of size $n + m$ from specific distribution $x_1, x_2, ..., x_n, x_{n+1}, ..., x_{n+m}$ to consider $x_1, x_2, ..., x_n$ is the past sample and $x_{n+1}, ..., x_{n+m}$ is the future sample.

2. Find the observed mean of m future observations $x_{n+1}, ..., x_{n+m}$, $\bar{x}_m$.

3. From each original sample draw $B = 1000$ NPI bootstrap samples of size m from $x_1, x_2, ..., x_n$ and calculate the mean of these values, $\bar{x}^*_m$. Now we have a list of B items of $\bar{x}^*_m$. Then construct $1 - \alpha$ prediction interval for the mean of future values: The lower bound is the $B.\frac{\alpha}{2}$th value in the ordered list of $\bar{x}^*_m$ in NPI bootstrap samples and the upper bound is the $B.(1 - \frac{\alpha}{2})$th value. Determine if this interval contains the mean of future observations in the original sample $\bar{x}_m$. If $B.\frac{\alpha}{2}$ or $B.(1 - \frac{\alpha}{2})$ are not integer, use the nearest integer.

4. After finishing this process for all the original samples we find the proportion of intervals which contain $\bar{x}_m$.

	$m = n$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	1	1	0.99	1	0.93	0.94	0.90	0.94	0.99
standard-B	0.93	0.96	0.95	0.90	0.93	0.78	0.79	0.82	0.80	0.85
	$m = 10$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.97	0.97	1	1	0.90	0.94	0.97	0.96	0.97
standard-B	0.96	0.95	0.97	0.99	1	0.79	0.91	0.95	0.96	0.95
	$m = 3$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	0.98	0.99	0.96	0.94	0.96	0.94	0.94
standard-B	0.99	0.99	1	0.98	0.99	0.93	0.93	0.95	0.94	0.93
	$m = 1$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.98	1	0.99	0.99	0.96	0.95	0.94	0.96	0.98
standard-B	0.85	0.96	0.99	0.99	0.99	0.91	0.93	0.93	0.97	0.97

Table 2.26: Coverage of $(1 - \alpha)$ prediction interval of some m observations from Uniform (0,1)

Repeat these steps with standard bootstrap samples and compare their performance. We constructed the prediction intervals of m future observations $m = n, 10, 3, 1$ by standard-B and NPI-B methods, as shown in the steps above. We study three distributions: Gamma (2,2), Uniform (0,1) and Normal (28,4) for various values of $n = 20, 50, 100, 200, 500$ and $\alpha = 0.01, 0.05$.

When $\alpha = 0.05$, $m = n$ and with Uniform (0,1), in Table 2.26, NPI-B has the largest coverage proportion, and the same results appear when $\alpha = 0.01$, but in this case the coverage proportion is closer to the nominal one. When $\alpha = 0.05$ and $m = 10$, both methods overcoverage when $n \geq 200$, but still the results of the NPI-B are higher than the standard-B. When $\alpha = 0.05$, there is no big difference between the two methods if $n \geq 100$, but when $n \leq 50$ the coverage proportion of NPI-B is larger than standard-B. If $m = 3$, the NPI-B and standard-B have closed results. However when $\alpha = 0.05$ the NPI-B has the largest results. When $\alpha = 0.05$ and $m = 1$, we can see that our method has a larger coverage proportion than the standard-B method, but sometimes overcoverage. When $\alpha = 0.01$, the two bootstrap methods achieve similar results when $n \geq 100$, but when $n = 20, 50$ the NPI-B achieves the best results.

	$m = n$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.98	1	0.99	0.98	0.98	0.94	0.98	0.97
standard-B	0.90	0.92	0.89	0.97	0.94	0.82	0.77	0.80	0.85	0.81
	$m = 10$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	0.99	1	1	0.95	0.96	0.96	0.97	0.97
standard-B	0.95	0.98	0.97	0.98	1	0.85	0.87	0.95	0.96	0.97
	$m = 3$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.97	0.98	0.99	0.99	0.92	0.95	0.97	0.97	0.98
standard-B	0.96	0.96	0.98	0.98	0.99	0.88	0.93	0.94	0.96	0.98
	$m = 1$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	0.99	1	0.95	0.88	0.91	0.99	0.99
standard-B	0.98	0.97	1	0.99	0.99	0.91	0.85	0.91	0.98	0.99

Table 2.27: Coverage of $(1 - \alpha)$ prediction interval of some m observations from Gamma (2,2)

With Gamma (2,2), in Table 2.27, and $\alpha = 0.01, 0.05$ and $m = n$ the coverage proportion of NPI-B is the largest, but when $\alpha = 0.05$ NPI-B shows overcoverage. Moreover the same occurred with $m = 10$. Note that we can use $(1-2\alpha)$ intervals as we did before with confidence intervals, but we need to change the lower and upper bounds to be $B.\alpha$th and $B(1-\alpha)$th values in the ordered list of $\bar{x}^*_m$, respectively. For the case $m = 3$ the overcoverage of NPI-B occurs when $n \geq 100$ and $\alpha = 0.05$, but it works well when $n = 20, 50$. When $\alpha = 0.01$, our method has a bigger coverage proportion than standard-B when $n = 20, 50$. Otherwise the two methods have similar results. For $m = 1$, we can see that the NPI-B is better than standard-B when $n = 20, 50$ if $\alpha = 0.01, 0.05$. In the other cases they are equivalent. The overcoverage of NPI-B appears when $\alpha = 0.05$ and $n \geq 100$.

Table 2.28, when $m = n$ and $\alpha = 0.01, 0.05$, shows that the NPI-B has good results because its coverage proportion is closer to the nominal coverage probability than standard-B. The two methods have similar results when $\alpha = 0.01$ and $m = 10$, except when n is small and the NPI-B works better than the standard-B. If $\alpha = 0.05$ and $m = 10$, our method has the largest coverage proportion, but sometimes overcoverage. The same status is shown in the next tables when $m = 3, 1$ for different values of α.

	$m = n$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.99	0.98	0.99	0.95	0.96	0.92	0.91	0.97
standard-B	0.85	0.95	0.94	0.94	0.95	0.80	0.82	0.78	0.82	0.84
	$m = 10$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	0.99	0.98	1	0.95	0.97	0.97	0.98	0.93
standard-B	0.95	0.99	0.99	0.98	0.99	0.83	0.93	0.95	0.98	0.92
	$m = 3$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.99	1	0.99	1	0.95	0.96	0.98	0.96	0.96
standard-B	0.95	0.98	0.97	0.99	0.99	0.86	0.90	0.97	0.95	0.96
	$m = 1$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.99	0.99	0.99	1	0.97	0.96	0.95	0.95	0.97
standard-B	0.89	0.94	0.98	0.99	1	0.94	0.93	0.92	0.95	0.98

Table 2.28: Coverage of $(1 - \alpha)$ prediction interval of some m observations from Normal (28,4)

Now, we want to repeat the process, but with a small difference. Here we will draw the past and the future samples seperately. If we draw the original sample of size n from the specific distribution $x_1, ..., x_n$ and then draw the future sample of size m from the same distribution $y_1, ..., y_m$ as happened in [60], we will refer to this by "alternative prediction interval" in order to distinguish between this method and the last method. We will compare the prediction intervals and the parameters of distribution of future samples. For example, if we have Normal (28,4) we will see if the prediction interval of the mean contains $\mu = 28$.

Table 2.29 shows that when $m = n$ and $\alpha = 0.01, 0.05$ the NPI-B has a better coverage than standard-B and it is closer to the proposed coverage probability (0.95 or 0.99). When $\alpha = 0.05$ and $m = 10$, the NPI-B has the largest coverage proportion but is still sometimes less than the nominal level 0.95 as appeared with $n = 20, 50$ or overcoverage when $n \geq 100$. Nevertheless, with $\alpha = 0.01$, the NPI-B and standard-B have similar coverage and close to nominal coverage probability. When $m = 3$ and $\alpha = 0.05$, the NPI-B has the largest coverage proportion, but is still sometimes less than the nominal level 0.95 as appeared with $n = 50, 200, 500$ or overcoverage when $n = 20, 100$. When $\alpha = 0.01$, the NPI-B and standard-B have similar coverage and close to nominal coverage probability except when $n = 20$,

	$m = n$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.99	0.98	0.99	1	0.93	0.95	0.96	0.95	0.93
standard-B	0.83	0.94	0.90	0.96	0.94	0.80	0.82	0.82	0.84	0.85
	$m = 10$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.98	0.99	1	0.98	0.90	0.94	0.97	0.96	0.97
standard-B	0.98	0.99	0.98	1	0.98	0.79	0.91	0.95	0.96	0.95
	$m = 3$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.96	1	0.99	1	0.97	0.96	0.94	0.96	0.94	0.94
standard-B	0.93	1	0.99	1	0.98	0.93	0.93	0.95	0.94	0.93
	$m = 1$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.99	1	1	0.99	0.94	0.93	0.96	0.94	0.96
standard-B	0.87	0.97	1	0.99	0.99	0.87	0.91	0.94	0.94	0.94

Table 2.29: Coverage of $(1-\alpha)$ alternative prediction interval of some m observations from Uniform (0,1)

NPI-B gets the largest value. Table 2.29 shows how, when $m = 1$ and $\alpha = 0.05$, the coverage of standard-B is less than NPI-B but is sometimes close to the nominal coverage probability as NPI-B. But if $\alpha = 0.01$, both methods have the nominal coverage except if $n = 20, 50$. In that situation the NPI-B is better than standard-B. In Table 2.30 with $m = n$, the NPI-B works well because it has the largest coverage proportions in all cases. The similar situation appears with $m = 10$. The NPI-B and standard-B for the case $m = 3$ have equivalent results in most cases. For $m = 1$ and $\alpha = 0.05$, the NPI-B has overcoverage proportion in most cases, but for $\alpha = 0.01$ it has better coverage than standard-B.

NPI-B with Normal (28,4) and $m = n$, $\alpha = 0.01, 0.05$, in Table 2.31, is the best because it has the highest coverage which is closer to the nominal coverage probability than standard-B. When $\alpha = 0.05$ and $m = 10$, the coverage of standard-B is close to the nominal coverage probability, whereas NPI-B overcoverage except when $n = 20, 50$. On the other hand, NPI-B is the best when $\alpha = 0.01$. When $m = 3$ and $\alpha = 0.05$ the Table 2.31 showed the overcoverage of NPI-B except when $n = 20$, whereas with $\alpha = 0.01$, it is desired the nominal coverage probability. When predicting a future observation from Normal (28,4), with $m = 1$, we can see that the NPI-B is the best when $\alpha = 0.01, 0.05$.

	$m = n$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	0.99	0.99	0.96	0.97	0.96	0.95	0.94
standard-B	0.90	0.91	0.97	0.90	0.92	0.71	0.84	0.78	0.82	0.83
	$m = 10$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.98	0.99	0.99	0.98	0.96	0.92	0.93	0.94	0.97
standard-B	0.95	0.97	0.97	0.99	0.98	0.88	0.86	0.89	0.90	0.98
	$m = 3$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	0.99	1	0.99	0.94	0.96	0.95	0.96	0.95
standard-B	0.96	0.99	0.99	1	0.99	0.88	0.93	0.94	0.95	0.94
	$m = 1$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.98	0.98	1	0.99	0.98	0.96	0.97	0.98	0.97
standard-B	0.88	0.95	0.96	0.99	0.98	0.94	0.93	0.93	0.98	0.97

Table 2.30: Coverage of $(1-\alpha)$ alternative prediction interval of some m observations from Gamma (2,2)

	$m = n$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.98	0.99	0.99	0.99	0.94	0.94	0.96	0.95	0.95
standard-B	0.88	0.89	0.94	0.93	0.95	0.81	0.84	0.85	0.83	0.85
	$m = 10$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.99	1	0.99	1	0.92	0.95	0.97	0.97	0.98
standard-B	0.92	0.97	0.99	0.98	1	0.87	0.91	0.94	0.94	0.97
	$m = 3$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	1	0.98	1	1	0.95	0.98	0.94	0.97	0.99
standard-B	0.92	0.98	0.95	1	1	0.93	0.96	0.93	0.96	0.99
	$m = 1$									
	$\alpha = 0.01$					$\alpha = 0.05$				
$n =$	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.97	0.98	1	0.96	0.91	0.92	0.94	0.95
standard-B	0.90	0.93	0.97	0.97	1	0.91	0.88	0.93	0.92	0.94

Table 2.31: Coverage of $(1-\alpha)$ alternative prediction interval of some m observations from Normal (28,4)

The summary of the above results, when $\alpha = 0.01$ and $m = n$, shows that the NPI-B is the best in most cases. If $m = 10, 3, 1$ we have two cases: if n is small, the NPI-B is the best and if n is large the two methods are equivalent. When $\alpha = 0.05$, if $m = n$ the NPI-B works better than standard-B. Additionally if $m = 10, 3, 1$ and n is small, but if n is large, the two methods are similar and overcoverage.

Mojirsheibani [60] and Mohirsheibani and Tibshirani [61] displayed different types of bootstrap prediction intervals to estimate the parameter θ: bootstrap-t, percentile and BCa prediction intervals, as discussed in detail in Section 1.3.4. These intervals are transformation respecting and range preserving. The range preserving property means that the procedure produced intervals that fall in the range of parameter. To construct the percentile prediction interval of a statistic as in [60, 61], we use this method by Mojirsheibani and Tibshirani, which we will refer to as the MT method:

1. Draw c original samples from any distribution, $x_1, x_2, ..., x_n$ to be the past sample and then draw $y_1, y_2, ..., y_m$ to be the future sample. X and Y are iid.

2. From each original sample find the statistic of the future sample, T_m, and then sample $B = 1000$ NPI-B samples (future samples) from $x_1, x_2, ..., x_n$ and find the statistic from each bootstrap samples T_m^* to get a list of T_{mj}^* where $j = 1, ..., B$.

3. Construct the $(1 - 2\alpha)$ percentile prediction interval of T_m:
   lower bound: $\hat{F}_m^{-1}[\Phi(z^{(\alpha)}(1 + \frac{m}{n})^{\frac{1}{2}})] = \hat{F}_m^{-1}[\alpha_1]$ is the $B.\alpha_1$th value of the ordered list of T_{mj}^*.
   upper bound: $\hat{F}_m^{-1}[\Phi(z^{(1-\alpha)}(1 + \frac{m}{n})^{\frac{1}{2}})] = \hat{F}_m^{-1}[\alpha_1]$ is the $B.\alpha_2$th value of the ordered list of T_{mj}^*. If $B.\alpha_1$ or $B.\alpha_2$ are not integer, use the largest integer.

4. Determine if this interval contains the statistic T_m and find the proportion of these intervals. For example, if the future sample is Normal (28,4) we will determine if the prediction interval of the mean contains the mean of Normal=28.

Lu and Chang [55] use the bootstrap method to construct prediction intervals for observations from the Birnbaum-Saunders distribution, which turned out to have good coverage results. They applied the bootstrap percentile method which we also apply below. We call it the classical method:

1. Draw an actual sample of size n from a specific distribution, giving $x_1, ..., x_n$. Then draw a second sample of size m from the same distribution, giving $y_1, ..., y_m$ with mean value $\bar{y}$. This will be used as the future sample to check the performance of the bootstrap prediction intervals.

2. Use the actual sample to draw B NPI-B samples of size m as described above. Calculate the mean of each NPI-B sample, giving m_j for $j = 1, \ldots, B$.

3. Construct an $100(1 - 2\alpha)\%$ prediction interval for the mean by defining the lower bound to be the $\alpha \times B$-th value in the ordered list of the values m_j and the upper bound to be the $(1 - \alpha) \times B$-th value in this list (use the largest integer if these values or not integer).

4. Check if the prediction interval from step 3 contains the mean $\bar{y}$ of the future sample from the underlying distribution as resulted from step 1.

This procedure is applied repeatedly to derive an indication of the coverage of these intervals, that is the proportion of such intervals which indeed contain the mean value of the future sample from the underlying distribution. For perfect coverage the probability of the $100(1 - 2\alpha)\%$ interval containing that mean value should of course be $100(1 - 2\alpha)\%$. This same procedure has also been used in the following section for the standard-B method, of course changing step 2 accordingly.

The Tables from 2.32 to 2.49 described how NPI-B has the largest values of coverage proportions but it overcoverage in most cases with MT method. The NPI-B performs well when the classical method is used, it is closer to the nominal coverage than the classical method.

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	0.99	1	1	0.94	0.97	0.96	0.99	0.99
standard-B	0.94	0.94	0.95	1	0.99	0.85	0.87	0.86	0.95	0.92
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	0.99	0.99	0.99	0.98	0.96	0.99	0.95	0.96
standard-B	0.93	0.96	0.98	0.97	0.96	0.85	0.88	0.85	0.88	0.89
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	1	1	0.98	0.98	0.99	0.99	0.99
standard-B	0.92	0.93	0.99	0.99	0.99	0.82	0.84	0.87	0.95	0.90

Table 2.32: Uniform (0,1), $m = n$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	1	0.99	0.99	0.95	0.92	0.93	0.95	0.91
standard-B	0.93	0.93	0.94	0.94	0.92	0.73	0.68	0.80	0.80	0.77
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.96	0.98	0.98	0.96	0.96	0.92	0.89	0.95	0.88
standard-B	0.88	0.91	0.88	0.94	0.85	0.82	0.70	0.78	0.80	0.69
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.98	1	0.99	0.93	0.88	0.92	0.92	0.89
standard-B	0.93	0.85	0.92	0.91	0.89	0.82	0.74	0.78	0.77	0.78

Table 2.33: Uniform (0,1), $m = n$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	0.99	1	0.98	0.98	0.99	0.98	1	0.98
standard-B	0.96	0.98	0.97	0.99	0.98	0.89	0.96	0.97	0.99	0.98
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	1	1	0.99	0.99	0.99	1	0.99	0.99	0.99
standard-B	0.97	0.99	1	0.97	0.99	0.93	0.98	0.98	0.97	0.99
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.98	0.99	0.99	0.97	0.97	0.98	0.99	0.99
standard-B	0.97	0.95	0.97	0.98	0.98	0.91	0.95	0.96	0.98	0.97

Table 2.34: Uniform (0,1), $m = 10$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.92	0.90	0.88	0.91	0.88	0.89	0.89	0.86	0.91	0.88
standard-B	0.86	0.89	0.87	0.92	0.87	0.80	0.87	0.83	0.90	0.87
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.95	0.96	0.91	0.93	0.93	0.92	0.93	0.91	0.92	0.93
standard-B	0.88	0.93	0.93	0.90	0.88	0.82	0.86	0.90	0.90	0.88
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.95	0.95	0.91	0.89	0.87	0.90	0.93	0.91	0.88	0.87
standard-B	0.88	0.93	0.87	0.88	0.88	0.83	0.87	0.85	0.88	0.88

Table 2.35: Uniform (0,1), $m = 10$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.92	0.91	0.90	0.91	0.89	0.90	0.90	0.89	0.91	0.89
standard-B	0.89	0.90	0.90	0.93	0.88	0.87	0.90	0.89	0.92	0.88
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.93	0.96	0.91	0.93	0.93	0.91	0.96	0.91	0.93	0.93
standard-B	0.99	0.95	0.90	0.92	0.93	0.88	0.93	0.88	0.92	0.93
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.89	0.92	0.95	0.92	0.90	0.89	0.91	0.95	0.92	0.90
standard-B	0.86	0.90	0.95	0.93	0.90	0.85	0.90	0.94	0.92	0.90

Table 2.36: Uniform (0,1), $m = 3$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.98	0.97	1	0.99	0.98	0.98	0.97	1
standard-B	0.97	0.99	0.95	0.97	1	0.95	0.98	0.95	0.97	1
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	1	0.97	0.99	0.96	0.99	1	0.97	0.99	0.96
standard-B	0.98	1	0.97	0.99	0.96	0.96	1	0.97	0.99	0.96
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.99	0.97	0.98	0.99	0.99	0.99	0.97	0.98
standard-B	0.95	0.98	0.97	0.98	0.99	0.94	0.97	0.96	0.98	0.99

Table 2.37: Uniform (0,1), $m = 3$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.94	0.98	0.98	0.93	0.87	0.82	0.91	0.92
standard-B	0.90	0.84	0.80	0.88	0.94	0.77	0.70	0.68	0.75	0.80
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.96	0.98	0.97	0.97	0.92	0.90	0.90	0.90	0.87
standard-B	0.87	0.88	0.86	0.91	0.85	0.75	0.66	0.68	0.71	0.69
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.95	0.98	0.98	0.97	0.94	0.89	0.86	0.90	0.85
standard-B	0.87	0.85	0.85	0.90	0.85	0.80	0.71	0.70	0.77	0.74

Table 2.38: Normal (28,4), $m = n$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	1	1	1	1	0.95	0.97	0.99	0.99
standard-B	0.96	0.94	0.96	0.98	0.98	0.92	0.88	0.89	0.93	0.93
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	1	1	0.98	1	0.99	0.99	1
standard-B	0.92	0.94	0.93	0.98	0.98	0.78	0.89	0.91	0.90	0.88
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	1	1	1	0.95	0.99	0.98	0.99
standard-B	1	0.95	0.97	0.99	0.99	0.94	0.83	0.87	0.90	0.92

Table 2.39: Normal (28,4), $m = n$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.94	0.97	0.95	0.94	0.89	0.89	0.92	0.94	0.94	0.89
standard-B	0.86	0.90	0.94	0.93	0.88	0.80	0.87	0.92	0.92	0.86
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.94	0.91	0.92	0.85	0.92	0.91	0.89	0.92	0.85
standard-B	0.87	0.84	0.87	0.89	0.83	0.74	0.79	0.86	0.89	0.83
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.95	0.93	0.93	0.90	0.88	0.93	0.93	0.92	0.90	0.88
standard-B	0.88	0.91	0.91	0.89	0.87	0.76	0.90	0.90	0.88	0.87

Table 2.40: Normal (28,4), $m = 10$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.93	0.92	0.94	0.89	0.86	0.93	0.91	0.94	0.87	0.86
standard-B	0.90	0.90	0.93	0.88	0.89	0.87	0.89	0.93	0.87	0.89
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	0.92	0.93	0.87	0.89	0.96	0.88	0.93	0.86	0.89
standard-B	0.94	0.84	0.90	0.83	0.90	0.88	0.84	0.90	0.83	0.90
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.96	0.89	0.93	0.90	0.89	0.95	0.87	0.92	0.90	0.89
standard-B	0.90	0.88	0.92	0.87	0.88	0.85	0.87	0.90	0.87	0.88

Table 2.41: Normal (28,4), $m = 3$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.99	0.98	0.96	0.96	0.98	0.99	0.98	0.96
standard-B	0.95	0.98	0.96	0.99	0.97	0.91	0.98	0.96	0.99	0.96
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	1	0.99	0.97	0.98	0.96	0.98	0.98	0.96	0.98
standard-B	0.90	0.94	0.97	0.96	0.98	0.84	0.93	0.96	0.95	0.98
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	1	1	0.99	0.99	0.97	0.99	0.99	0.99	0.98
standard-B	0.95	0.99	1	0.98	0.98	0.91	0.97	0.98	0.98	0.98

Table 2.42: Normal (28,4), $m = 10$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.95	0.98	0.98	0.98	0.98	0.94	0.98	0.98	0.98
standard-B	0.96	0.92	0.96	0.98	0.97	0.95	0.92	0.96	0.98	0.97
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.98	0.98	1	0.99	0.97	0.98	0.98	1	0.99
standard-B	0.94	0.97	0.97	1	0.98	0.94	0.97	0.97	1	0.98
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.98	0.98	0.98	0.99	0.98	0.97	0.97	0.98	0.99
standard-B	0.94	0.93	0.97	0.98	0.97	0.92	0.92	0.97	0.98	0.97

Table 2.43: Normal (28,4), $m = 3$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	1	1	0.99	0.98	1	0.98	0.99
standard-B	0.95	0.97	0.97	0.96	0.99	0.86	0.89	0.92	0.91	0.94
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	1	1	0.99	1	0.97	1	0.99	0.98
standard-B	0.89	0.90	0.94	0.98	0.96	0.80	0.86	0.85	0.90	0.92
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	1	1	0.99	0.99	0.98	0.97	0.99
standard-B	0.96	0.97	0.97	0.97	0.99	0.89	0.91	0.87	0.91	0.91

Table 2.44: Gamma (5,2), $m = n$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	0.99	0.97	0.97	1	1	0.99	0.97	0.97
standard-B	0.96	0.98	0.99	0.96	0.98	0.90	0.97	0.99	0.96	0.98
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	0.97	1	0.98	1	0.99	0.97	1	0.98
standard-B	0.89	0.96	0.92	0.98	0.98	0.84	0.93	0.90	0.98	0.98
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	1	1	0.97	0.98	0.99	0.98	1	0.96	0.98
standard-B	0.96	0.99	0.97	0.97	0.99	0.93	0.98	0.96	0.96	0.99

Table 2.45: Gamma (5,2), $m = 10$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	0.98	0.98	0.98	0.99	0.99	0.98	0.97	0.97
standard-B	0.97	0.99	0.98	0.92	0.99	0.96	0.99	0.98	0.92	0.99
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	1	0.99	0.98	0.99	0.99	1	0.99	0.98	0.99	0.98
standard-B	0.95	0.98	0.98	0.98	0.97	0.94	0.98	0.98	0.98	0.97
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.98	0.99	0.99	0.96	0.99	0.98	0.99	0.98	0.96
standard-B	0.95	0.98	0.97	0.97	0.96	0.94	0.98	0.97	0.95	0.96

Table 2.46: Gamma (5,2), $m = 3$, $\alpha = 0.01$, 0.98 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.96	0.91	0.90	0.96	0.88	0.93	0.90	0.90	0.96	0.88
standard-B	0.85	0.89	0.88	0.96	0.88	0.82	0.89	0.88	0.96	0.88
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.93	0.91	0.92	0.91	0.92	0.93	0.91	0.90	0.91	0.92
standard-B	0.84	0.88	0.89	0.90	0.89	0.81	0.88	0.89	0.89	0.89
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.95	0.90	0.90	0.94	0.88	0.93	0.88	0.89	0.94	0.88
standard-B	0.87	0.85	0.89	0.93	0.88	0.82	0.85	0.87	0.93	0.88

Table 2.47: Gamma (5,2), $m = 3$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.95	0.93	0.97	0.87	0.93	0.94	0.91	0.96	0.86
standard-B	0.91	0.89	0.89	0.95	0.85	0.81	0.87	0.88	0.94	0.85
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.91	0.95	0.96	0.92	0.94	0.89	0.93	0.96	0.92
standard-B	0.78	0.79	0.91	0.92	0.90	0.72	0.78	0.89	0.90	0.90
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.92	0.90	0.94	0.86	0.95	0.91	0.89	0.94	0.86
standard-B	0.90	0.91	0.88	0.92	0.85	0.83	0.87	0.85	0.92	0.85

Table 2.48: Gamma (5,2), $m = 10$, $\alpha = 0.05$, 0.90 prediction interval

MT method						classical method				
mean										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.98	0.99	0.99	0.96	0.93	0.92	0.93	0.93	0.79	0.84
standard-B	0.85	0.91	0.86	0.78	0.83	0.76	0.82	0.81	0.65	0.72
variance										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.99	0.99	0.96	1	0.95	0.94	0.94	0.93	0.89	0.85
standard-B	0.79	0.88	0.86	0.83	0.88	0.64	0.71	0.74	0.70	0.73
q_{75}										
n=m	20	50	100	200	500	20	50	100	200	500
NPI-B	0.97	1	0.98	0.98	0.97	0.91	0.92	0.91	0.91	0.89
standard-B	0.89	0.91	0.91	0.92	0.86	0.83	0.79	0.74	0.76	0.71

Table 2.49: Gamma (5,2), $m = n$, $\alpha = 0.05$, 0.90 prediction interval

2.6 NPI-B for Order Statistics

We studied the performance of the NPI-B method using confidence intervals the prediction intervals. Now we want to discuss the performance of the NPI-B method with order statistics. In [23] NPI for order statistics of m future observations and the way it was used to compare two groups was presented. Let $X_{(1)}, X_{(2)}, ..., X_{(m)}$ be the ordered statistics of m future observations and $X_{(r)}$, $r = 1, ..., m$, be the r-th ordered future observation. Coolen and Maturi [23] derived the following probability:

$$P(X_{(r)} \in I_j) = \binom{j+r-2}{j-1}\binom{n-j+1+m-r}{n-j+1}\binom{n+m}{n}^{-1} \tag{2.8}$$

to find the probability that r-th ordered future statistic $X_{(r)}$ belongs to interval I_j, where is $j = 1, ..., n+1$ and $r = 1, ..., m$. They also consider the limit results of this probability if m goes to infinity. To compare the two groups X and Y based on the r-th future order statistics of these groups they considered the event $X_{(r)} < Y_{(r)}$ and found its NPI lower and upper probabilities. As discussed before, the NPI method is based on the same $A_{(n)}$ assumption as the NPI-B approach, so we show a small example to discuss how NPI-B achieves the formula (2.8). If the original sample $X = (0.294, 0.801, 0.971, 0.987)$ is drawn from Uniform (0,1), then we sample 1000 NPI-B samples of size $m = n = 4$ from X and study the ordered statistics $X_{(2)}$ and $X_{(4)}$. Tables 2.50 and 2.51 show the probabilities of each order statistic belonging to the indicated interval.

$X_{(2)}$		
Probabilities	results from formula (2.8)	results from simulation
$P(X_{(2)} \in I_1)$	0.20	0.22
$P(X_{(2)} \in I_2)$	0.27	0.28
$P(X_{(2)} \in I_3)$	0.26	0.24
$P(X_{(2)} \in I_4)$	0.22	0.20
$P(X_{(2)} \in I_5)$	0.05	0.07

Table 2.50: NPI-B with $X_{(2)}$ order statistics, $n = m = 4$

$X_{(4)}$		
Probabilities	results from formula (2.8)	results from simulation
$P(X_{(4)} \in I_1)$	0.01	0.02
$P(X_{(4)} \in I_2)$	0.05	0.05
$P(X_{(4)} \in I_3)$	0.17	0.15
$P(X_{(4)} \in I_4)$	0.28	0.30
$P(X_{(4)} \in I_5)$	0.49	0.49

Table 2.51: NPI-B with $X_{(4)}$ order statistics, $n = m = 4$

$X_{(2)}$					
Probabilities	results from formula (2.8)	results from simulation	Probabilities	results from formula (2.8)	results from simulation
$P(X_{(2)} \in I_1)$	0.24	0.23	$P(X_{(2)} \in I_{12})$	4×10^{-4}	1×10^{-3}
$P(X_{(2)} \in I_2)$	0.26	0.27	$P(X_{(2)} \in I_{13})$	1×10^{-4}	0
$P(X_{(2)} \in I_3)$	0.20	0.21	$P(X_{(2)} \in I_{14})$	4×10^{-5}	0
$P(X_{(2)} \in I_4)$	0.13	0.13	$P(X_{(2)} \in I_{15})$	1×10^{-5}	0
$P(X_{(2)} \in I_5)$	0.08	0.07	$P(X_{(2)} \in I_{16})$	3×10^{-6}	0
$P(X_{(2)} \in I_6)$	0.05	0.05	$P(X_{(2)} \in I_{17})$	9×10^{-7}	0
$P(X_{(2)} \in I_7)$	0.02	0.02	$P(X_{(2)} \in I_{18})$	1×10^{-7}	0
$P(X_{(2)} \in I_8)$	0.01	0.01	$P(X_{(2)} \in I_{19})$	2×10^{-8}	0
$P(X_{(2)} \in I_9)$	0.01	0.01	$P(X_{(2)} \in I_{20})$	2×10^{-9}	0
$P(X_{(2)} \in I_{10})$	2×10^{-3}	4×10^{-3}	$P(X_{(2)} \in I_{21})$	1×10^{-10}	0
$P(X_{(2)} \in I_{11})$	1×10^{-3}	0			

Table 2.52: NPI-B with $X_{(2)}$ order statistics, $n = m = 20$

The first column in each table shows the probabilities using the formula (2.8), and the second shows these probabilities using simulated NPI-B samples from X. For example to find $P(X_{(2)} \in I_1)$, count the NPI-B samples which have $X_{(2)}$ in the first interval. Note that $I_1 = (0, 0.294)$, $I_2 = (0.294, 0.801)$,..., and $I_{n+1} = (0.987, 1)$. By exploring the values of two probabilities in Tables 2.50 and 2.51, we see that these values are close in most cases, and that means the NPI-B method works in line with the formula. When we consider the same idea but with a larger sample size, we use the original sample from Uniform (0,1) and $n = m = 20$ and study the same order statistics. The results are explored in Tables 2.52 and 2.53. The larger sample size does not improve the agreement between the threoretical and actual probabilities of $X_{(2)}$ as much, but with $X_{(4)}$ the situation is different, and the agreement is clearly improves.

$X_{(4)}$					
Probabilities	results from formula (2.8)	results from simulation	Probabilities	results from formula (2.8)	results from simulation
$P(X_{(4)} \in I_1)$	0.05	0.06	$P(X_{(4)} \in I_{12})$	0.01	0.01
$P(X_{(4)} \in I_2)$	0.12	0.11	$P(X_{(4)} \in I_{13})$	2×10^{-3}	0
$P(X_{(4)} \in I_3)$	0.16	0.17	$P(X_{(4)} \in I_{14})$	9×10^{-4}	2×10^{-3}
$P(X_{(4)} \in I_4)$	0.17	0.16	$P(X_{(4)} \in I_{15})$	3×10^{-4}	1×10^{-3}
$P(X_{(4)} \in I_5)$	0.15	0.15	$P(X_{(4)} \in I_{16})$	1×10^{-4}	0
$P(X_{(4)} \in I_6)$	0.12	0.12	$P(X_{(4)} \in I_{17})$	3×10^{-5}	0
$P(X_{(4)} \in I_7)$	0.09	0.09	$P(X_{(4)} \in I_{18})$	8×10^{-6}	0
$P(X_{(4)} \in I_8)$	0.06	0.06	$P(X_{(4)} \in I_{19})$	1×10^{-6}	0
$P(X_{(4)} \in I_9)$	0.04	0.04	$P(X_{(4)} \in I_{20})$	1×10^{-7}	0
$P(X_{(4)} \in I_{10})$	0.02	0.03	$P(X_{(4)} \in I_{21})$	1×10^{-8}	0
$P(X_{(4)} \in I_{11})$	0.01	0.01			

Table 2.53: NPI-B with $X_{(4)}$ order statistics, $n = m = 20$

2.7 Concluding Remarks

In this chapter the NPI-B method has been presented for distributions with finite and infinite intervals. First, we studied the performance of NPI-B by calculating the variance of statistics, bias, absolute error and MSE. It can be seen that all these methods have underestimate variance, but the NPI bootstrap does best because the variance of statistics in the NPI bootstrap is closer to the variance of the original sample than other methods. So we avoid underestimating the variance in the standard bootstrap because in our method we have more variations due to adding the observations to the data. Also, the NPI-B method goes outside the range of values in the standard bootstrap. The data show the values of MSE in the NPI bootstrap method are large, even in cases which have a small bias, because the NPI bootstrap method has a larger variance than other methods. This is natural, because it resamples the observations from the original sample and from the intervals between them, but the standard bootstrap resamples from the original sample only. We can see that the NPI bootstrap method performs better than other methods, for bias, in estimating variance.

When we looked into the NPI bootstrap method in other situations (real line quantities and non negative observations), with different distributions such as Gamma, Weibull, Normal and Students't-distribution, and by comparing standard and NPI bootstrap, we noted that the bias of NPI bootstrap was in most cases smaller than the bias of standard bootstrap, but MSE and absolute error were the largest, except

in few cases. The results showed that the two methods have underestimate variance of statistics, but NPI bootstrap had variances closer to the variance of the underlying population. That is one of the benefit of our method. There was one case had overestimate variance of statistic in NPI bootstrap, when we estimated the variance parameter from Gamma (2,2). Here we showed what the original distribution was, in order to explain the performance of the NPI-B method with various shapes of distributions.

When some difficulties appeared with measurements of point estimation, we decided to work with confidence intervals and prediction intervals. We found that NPI-B did not work well in estimation with confidence intervals. It worked well when we used different prediction intervals, because it has the best coverage probability in most cases in our study. The simulation in this chapter shows that it is a promising alternative to the standard bootstrap for prediction. The variability in the predictions reflects the variability in the underlying population, which is taken into account in the simulation study by the fact that the future sample was taken from the same underlying population as the actual data.

Finally, we considered the NPI-B method with order statistics, using results from the literature that provided a probability of r-th order statistic using the NPI aspect, and found that NPI-B samples were consistent with those results. That can be natural because NPI and NPI-B depend on the same assumption $A_{(n)}$. When we used a larger sample size, this consistency developed with different levels of order statistics. However this is a small study which gives a picture of the use of order statistics with NPI-B.

Chapter 3

NPI for Reproducibility of Basic Tests

3.1 Introduction

In this Chapter the NPI approach is presented for reproducibility probability (RP) for some basic nonparametric tests. The RP for a test is the probability for the event that, if the test is repeated based on an experiment performed in the same way as the original experiment, the test outcome, that is either the rejection or non rejection of the null hypothesis, will be the same. The importance of the RP of tests and some definitions of it were discussed in [43, 44, 59, 64]. It is used in cases where evidence in clinical trials is often strongly in favour of a new treatment [9, 65]. Also,the estimated RP is used to define tests, which provides an interesting alternative to tests based mainly on chosen significance tests [32–34]. The review of the literature on this topic was presented in Section 1.4.

In Section 3.2 we briefly review some basic nonparametric tests. The use of the NPI approach to RP with those tests is considered in Sections 3.3, 3.4, 3.5 and 3.6. This chapter finishes with some concluding remarks in Section 3.7.

3.2 Overview of Some Basic Tests

In this section we illustrate an overview of some nonparametric tests: one sample Sign test, one sample signed rank test and two sample rank sum test. These tests will be used in this Chapter and in Chapter 4 to explore the NPI and NPI-B methods with reproducibility probability.

3.2.1 One Sample Sign Test

Perhaps the most basic nonparametric test is the sign test [49, 52, 67]. Suppose we have n real valued random quantities $X_1, X_2, ..., X_n$, which are traditionally assumed to be mutually independent and identically distributed with median m_0, so $P(X_i < m_0) = P(X_i > m_0) = 1/2$ for $i = 1, ..., n$. Generally, we test the hypotheses

$$H_0 : \theta = m_0 \text{ versus } H_1 : \theta \neq m_0, > m_0, < m_0 \tag{3.1}$$

This test assumes that the data is iid from a continuous distribution with a positive density. The test statistic K is the number of these X_i that are positive, so

$$K = \sum_{i=1}^{n} I\{X_i > 0\} \tag{3.2}$$

with indicator function $I\{A\} = 1$ if A is true and $I\{A\} = 0$ if A is not true, and ignoring the observations which are equal m_0. For the one sided upper tail test with the level of significance α, $H_1 : \theta > m_0$, we reject H_0 if $K \geq b_{\alpha,1/2}$ with $b_{\alpha,1/2}$ the upper α percentile point for the Binomial distribution with sample size n and success probability $p = 1/2$, while reject H_0 if $K \leq n - b_{\alpha,1/2}$ for one sided lower tail test $H_1 : \theta < m_0$, and for two sided test $H_1 : \theta \neq m_0$ reject H_0 if $K \geq b_{\alpha/2,1/2}$ or $K \leq n - b_{\alpha/2,1/2}$. Where $b_{\alpha,1/2}$, $b_{\alpha/2,1/2}$ are given in some literature tables to make the type 1 error probability equal to α. If $n \to \infty$ we use standard normal distribution as an approximation with $\mu_K = \frac{n}{2}$ and $\sigma_K^2 = \frac{n}{4}$, the standardized version of K is K^* is:

$$K^* = \frac{K - \mu_K}{\sigma_K} = \frac{(K + 0.5) - 0.5 * n}{\frac{\sqrt{n}}{2}} \tag{3.3}$$

We reject H_0 if $K^* \geq z_\alpha$ for a one sided upper tail test and if $K^* \leq -z_\alpha$ for a lower tail test and reject H_0 if $|K^*| \geq z_{\alpha/2}$ for a two sided test.

3.2.2 One Sample Signed Rank Test

The one sample Wilcoxon Signed Rank test (WRS) [49, 52, 67, 69] is an improvement to the sign test if the population is symmetric about the median m_0. It is a popular nonparametric location test which takes more information from the sample into account than the sign test. Details about the history of the signed rank test and the corresponding standard frequentist theory, together with tables for critical values for the test statistic and approximations for large samples, can be found in many statistics textbooks, e.g. [41, 49]. Let $X_1, X_2, ..., X_n$ is an independent sample from an absolutely continuous, symmetric distribution, then the test statistic used in this test is:

$$W = \sum_{X_i > m_0} rank(|X_i - m_0|) \tag{3.4}$$

where rank $(|X_i - m_0|)$ is the rank of $|X_i - m_0|$, so the test statistic is the sum of ranks of such absolute differences for observations that are greater than the median. The assumption of an absolutely continuous underlying distribution is for convenience, as it reduces the requirement for dealing with ties. If there are ties in the absolute differences in the data these can be dealt with [41]. For the one sided upper tail test $H_0 : \theta = m_0$ versus $H_1 : \theta > 0$, we reject H_0 if $W \geq W_\alpha$, where W_α is the critical value for the test statistic for significance level α, and reject H_0 if $W \leq \frac{n(n+1)}{2} - W_\alpha$ if we have a one sided lower tail test with $H_1 : \theta < m_0$. If we use the two sided test $H_1 : \theta \neq m_0$, we reject H_0 if $W \geq W_{\alpha/2}$ or $W \leq \frac{n(n+1)}{2} - W_{\alpha/2}$. If $n \to \infty$, use $W^* = \frac{W - \mu_w}{\sigma_w}$ is $N(0,1)$ where $\mu_w = \frac{n(n+1)}{4}$ and $\sigma_w = \sqrt{\frac{n(n+1)(2n+1)}{24}}$. Reject H_0 if $W^* \geq z_\alpha$ for the upper tail test, if $W^* \leq -z_\alpha$ for the lower tail test and if $|W^*| \geq z_{\alpha/2}$ for the two sided test.

For ease of presentation we will assume in this thesis that there are no ties in the data. Adapting the NPI-RP approach, which is shown in the next section, for such possible ties is relatively easy, e.g. by breaking the ties in all possible ways and

taking the most conservative corresponding lower and upper probabilities. However, this is not of major practical relevance and is not discussed further here.

3.2.3 Two Sample Rank Sum Test

The comparison of two samples is one of the most common applications of statistical methods, with the two sample rank sum test, also known as the Wilcoxon Mann Whitney test (WMT) the most popular non parametric test for such scenarios [41,49]. For this test, data $X_1, X_2, ..., X_{n_1}$ are assumed to be an independent and identically distributed sample from a population with the cumulative probability distribution F, and data $Y_1, Y_2, ..., Y_{n_2}$ an independent and identically distributed sample from a population with cumulative probability distribution G, where also the X and Y observations are mutually independent. The two sample rank sum test considers null hypothesis

$$H_0 : F(t) = G(t), \text{ for all real valued } t \tag{3.5}$$

The null hypothesis asserts that the X variable and Y variable have the same probability distribution, but the common distribution is not specified. The alternative hypothesis specified that Y is larger (or smaller) than X. The model which describes the alternative is called the location shift model

$$H_1 : G(t) = F(t - \delta), \text{ for all } t \tag{3.6}$$

This means that the population 2 is the same as population 1 except that it is shifted by the amount δ. It can be written as

$$Y =^d X + \delta \tag{3.7}$$

and the null hypothesis can be written as

$$H_0 : \delta = 0 \tag{3.8}$$

and the usual alternative hypotheses are either two sided, that is $\delta \neq 0$ or one sided, so either $\delta > 0$ or $\delta < 0$. In this thesis we restrict attention to the one sided

upper tail alternative $H_1 : \delta > 0$. The two sided test involves more complicated combinatorics and is left as a possible topic for future research. As throughout this thesis, we assume that there are no ties in the data set in order to avoid making the presentation more complicated than needed to get the main point of this work across, namely the possibility of using NPI for inference on reproducibility of tests. If the mean of population 1 is $E(X)$ and the mean of population 2 is $E(Y)$, then

$$\delta = E(Y) - E(X) \tag{3.9}$$

To compute the Wilcoxon two sample rank sum test statistic Z, we order the combined sample of X and Y from small to large values. Let S_1 be the rank of Y_1, S_2 is the rank of Y_2 ... and S_n is the rank of Y_{n_2}. Let V_j be the rank assigned to Y_j and define the rank sum

$$Z = \sum_{j=1}^{n_2} V_j \tag{3.10}$$

The one upper sided test is

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta > 0 \tag{3.11}$$

reject H_0 if $Z \geq Z_\alpha$, with the critical value Z_α such that under the null hypothesis, $P(Z \geq Z_\alpha)$ for the chosen level of significance α, it is use the sum of ranks of the smaller sample size. The values of Z_α are typically provided in tables [41, 49].

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta < 0 \tag{3.12}$$

and reject H_0 if $Z \leq n_2(n_1 + n_2 + 1) - Z_\alpha$.
The two sided test is

$$H_0 : \delta = 0 \text{ versus } H_1 : \delta \neq 0 \tag{3.13}$$

and we reject H_0 if $Z \geq w_{\frac{\alpha}{2}}$ or $Z \leq n_2(n_1+n_2+1) - Z_{\frac{\alpha}{2}}$. Note that the R code gives the U statistic instead of the Z statistic,which is called Mann Whitney U statistic

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \phi(X_i, Y_j) \tag{3.14}$$

and $\phi(X_i, Y_j) = 1$ if $X_i < Y_j$ and 0 otherwise

$$Z = U + \frac{n_2(n_2+1)}{2} \tag{3.15}$$

that means that tests based on U are equivalent to tests based on Z. If n is large, we use the standard normal distribution as approximation

$$Z^* = \frac{Z - \mu_z}{\sigma_z} \tag{3.16}$$

where is $\mu_z = \frac{n_1(n_1+n_2+1)}{2}$ and $\sigma_z = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$.

3.3 NPI for the Reproducibility Probability

The reproducibility of a test is an important characteristic of the practical relevance of test outcomes. Recently there has been substantial interest in the reproducibility probability (RP), where not only its estimation but also it is the actual definition and interpretation are not uniquely determined in the classical frequentist statistics framework. NPI is a frequentist statistics approach that makes few assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty, and which explicitly focuses on future observations. The explicitly predictive nature of NPI provides a natural formulation of inferences on RP.

In the Sections 3.4, 3.5 and 3.6, we introduce the use of the NPI approach to RP (NPI-RP) for some basic nonparametric tests [20]. Applying NPI, for either real-valued or Bernoulli data, enables inference by deriving lower and upper probabilities for the event that a future test, of similar size and under similar circumstances as the first test, will lead to the same conclusion as the first test, that is rejection or non-rejection of the null-hypothesis. Generally, we will use the acronym NPI-RP for such inferences. It is important to emphasize that we focus on the conclusion of the future test with regard to the null-hypothesis, given the actual data of the first test; so we do not consider an exact repetition in terms of the same value for the test statistic of interest or even for the actual observations, nor do we opt to just use the information from the first test that the null-hypothesis was rejected or not. As the

strength of the first test's conclusion depends on the actual data, it seems logical and important to use those data to infer on the reproducibility of the test result, while such prediction for the test result in a future test is more naturally reflected by the corresponding final conclusion, so rejection or not of the null-hypothesis. In Section 3.7 we briefly comment on the possibility, within the NPI framework, to only use the test's conclusion from the first test, but we consider this of less importance than the approach followed throughout this chapter.

As is clear from the brief comments on the literature on RP in Section 1.4, there have been several different formulations of the RP problem within the classical theory of frequentist statistics, where typically properties of an assumed underlying population are estimated. However, the very nature of RP seems to be predictive; given the data from the first test, one would like to predict the overall test conclusion for a second test, if such a further test would have the same sample size(s) and would be performed under similar circumstances. Hence the NPI approach is attractive, as it is a framework of frequentist statistics that explicitly considers future observations which are exchangeable with the available data observations. We should point out that the NPI framework does not require that the sample size(s) in the actual (first) and future (second) tests are the same, but this seems a natural assumption in order to reflect reproducibility, and we will restrict attention to this situation in this thesis.

We present NPI-RP for any possible results of the first test, so both in case that it leads to rejection and non-rejection of the null-hypothesis. As will be clear from the discussion in Section 1.4, in practice one is often particularly interested in reproducibility of tests that led to rejection of the null-hypothesis, as this tends to be the practically most important scenario, e.g. leading to new medication being introduced. However, for a complete view we believe that the reproducibility of tests that did not reveal a significant effect is also important, so while our discussions (including in the examples in this thesis) will mostly focus on the reproducibility of tests in cases where the null-hypothesis is rejected, we also consider RP in cases of non-rejection of the null-hypothesis. The NPI for Bernoulli observations is used for the NPI approach to reproducibility for the sign test, presented in Section 3.4. The NPI for multiple real-valued observations is used for the NPI approach to the

reproducibility of the one-sample signed-rank test and the two-sample rank sum test, presented in Sections 3.5 and 3.6.

Before we consider NPI-RP for the sign test, which is perhaps the most basic nonparametric statistical test, we need to comment briefly on assumptions underlying statistical tests. Generally, when a statistical test is applied there are some modelling assumptions which, ideally, should be checked. For example, Wilcoxon's one-sample signed-rank test, which we consider in Section 3.5, assumes that the population from which the sample is drawn is symmetric about the median. This assumption is important for the distribution of the test statistic under the null-hypothesis and ideally should be checked whenever the test is applied.

In the NPI-RP approach, given the n data observations from the first test, we consider all possible different orderings of n future observations and the n data observations for this test, and then calculate lower and upper probabilities for the event that the test statistic based on such n future observations will lead to rejection or non-rejection of the null-hypothesis. When doing so, one could argue that we should consider, for each of the $\binom{2n}{n}$ possible orderings of the n future observations among the n data observations, whether or not it is reasonable to assume that the n future observations could have come from a population that is symmetric about its median. While this could be done, e.g. by using an appropriate pre-test, we do not do this for three reasons.

First, we will typically consider quite small data sets (although the approach can be applied for all sample sizes), in which case for only few test results such an underlying assumption would be rejected when formulated as null-hypothesis for a pre-test. Secondly, implementing such a pre-test for the predicted future samples would severely complicate both computation and analytic derivation of the results presented in this chapter. Thirdly, and most importantly, while testing such assumptions, or at least good awareness of such assumptions, it is indeed important for the actual (first) test, the further tests as performed on all the predictive, and hence hypothetical, future data sets are mainly done to get an insight in the corresponding values of the test statistic and the corresponding test conclusions; as we do not base the practically important overall conclusion on a single test out of

these predictive tests, whether or not the predicted data actually would support the underlying assumption is of less relevance. So, generally, we do not consider such underlying assumptions in this chapter, but we will assume that the method is only applied where such assumptions seem reasonable for the actual data from the first test, as is common when such tests are applied.

3.4 NPI-RP for the One Sample Sign Test

The NPI-RP approach for this scenario is as follows. Suppose that $Y = y$, then this test immediately leads to H_0 being rejected or not. We then consider NPI for the random number of positive observations out of n future Bernoulli observations [16], which we denote by Y_f (used for simplicity instead of the notation Y_{n+1}^{2n} which would be in line with notation used in Section 1.2; we only consider the situation with future number of observations equal to the number of data observations, so $m = n$, henceforth in this thesis), we discuss this briefly in Section 3.7, given the y positive observations out of n in the original test. So here we use equations (1.4) and (1.5). We derive the NPI lower and upper probabilities for the event that, for these n future observations, the test's conclusion will be the same as the original conclusion based on the observation y. This provides the NPI lower and upper reproducibility probabilities, which we denote by $\underline{RP}(y)$ and $\overline{RP}(y)$, respectively, where it is important to emphasize that these depend on y.

We first consider the one-sided test with $H_1 : \theta > 0$, for which H_0 is rejected if and only if $Y \geq b_\alpha$. The NPI lower and upper probabilities for reproducibility of this test involve consideration of the event $Y_f \geq b_\alpha$, given data $Y = y$ from the original test. The NPI lower and upper probabilities for this event are derived from (1.4) and (1.5), which are based on the assumption $A_{(n)}$ and the model presented

by Coolen [16], and are equal to

$$\underline{P}(Y_f \geq b_\alpha | y) =$$
$$1 - \binom{2n}{n}^{-1} \times \left[\binom{2n-y}{n-y} + \sum_{l=1}^{b_\alpha - 1} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]$$
$$\overline{P}(Y_f \geq b_\alpha | y) =$$
$$\binom{2n}{n}^{-1} \times \left[\binom{y+b_\alpha}{y} \binom{2n-y-b_\alpha}{n-y} + \sum_{l=b_\alpha+1}^{n} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]$$

The NPI lower and upper reproducibility probabilities in this case are as follows. For $y \geq b_\alpha$, so in case the original test led to rejection of H_0, $\underline{RP}(y) = \underline{P}(Y_f \geq b_\alpha | y)$ and $\overline{RP}(y) = \overline{P}(Y_f \geq b_\alpha | y)$, while for $y < b_\alpha$, which led to H_0 not being rejected in the original test, $\underline{RP}(y) = \underline{P}(Y_f < b_\alpha | y) = 1 - \overline{P}(Y_f \geq b_\alpha | y)$ and $\overline{RP}(y) = \overline{P}(Y_f < b_\alpha | y) = 1 - \underline{P}(Y_f \geq b_\alpha | y)$.

For the one-sided test with $H_1 : \theta < 0$, for which H_0 is rejected if and only if $Y \leq n - b_\alpha = b_\alpha^l$, the relevant NPI lower and upper probabilities, given $Y = y$, are also easily derived from (1.4) and (1.5) and are equal to

$$\underline{P}(Y_f \leq b_\alpha^l | y) = 1 - \binom{2n}{n}^{-1} \times$$
$$\left[\binom{y+b_\alpha^l+1}{y} \binom{2n-y-b_\alpha^l-1}{n-y} + \sum_{l=b_\alpha^l+2}^{n} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]$$
$$\overline{P}(Y_f \leq b_\alpha^l | y) = \binom{2n}{n}^{-1} \times \left[\binom{2n-y}{n-y} + \sum_{l=1}^{b_\alpha^l} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]$$

The NPI lower and upper reproducibility probabilities in this case are as follows. For $y \leq b_\alpha^l$, so in case the original test led to rejection of H_0, $\underline{RP}(y) = \underline{P}(Y_f \leq b_\alpha^l | y)$ and $\overline{RP}(y) = \overline{P}(Y_f \leq b_\alpha^l | y)$, while for $y > b_\alpha$, which led to H_0 not being rejected in the original test, $\underline{RP}(y) = \underline{P}(Y_f > b_\alpha^l | y) = 1 - \overline{P}(Y_f \leq b_\alpha^l | y)$ and $\overline{RP}(y) = \overline{P}(Y_f > b_\alpha^l | y) = 1 - \underline{P}(Y_f \leq b_\alpha^l | y)$.

For this one-sided sign test with $H_1 : \theta > 0$, the minimum value that can occur for the NPI lower reproducibility probability is equal to $\underline{RP}(b_\alpha) = \underline{P}(Y_f \geq b_\alpha | y = b_\alpha) = 0.5$, while with $H_1 : \theta < 0$ also the minimum value that can occur for the NPI lower reproducibility probability is equal to $\underline{RP}(b_\alpha^l) = \underline{P}(Y_f \leq b_\alpha^l | y = b_\alpha^l) = 0.5$ (the

justification for these values is given in Proof of minimum value for lower RP in this section). As these NPI lower reproducibility probabilities $\underline{RP}(y)$ are increasing in y, this minimum possible value of 0.5 links nicely to the discussions in the literature about reproducibility probability of 0.5 as a worst-case scenario. Note that in the NPI approach, this worst-case scenario is reflected through the lower probability, the corresponding NPI upper probability $\overline{RP}(b_\alpha)$ is greater than 0.5 and depends on n, in fact it is a decreasing function of n with limiting value 0.5 for $n \to \infty$ (this is easily proven but is left to the interested reader, it is of little further relevance in this thesis).

The maximum value that can occur for the NPI upper reproducibility probability for these tests is equal to 1, which occurs if all observations in the original test sample are positive, or if all are negative. In both these cases, these data observations are maximally supportive of either the null-hypothesis or the alternative hypothesis. These NPI upper probabilities reflect that, if all observations are positive (negative) then the data do not provide evidence against the possibility that negative (positive) observations would never occur.

For the two-sided test with $H_1 : \theta \neq 0$, for which H_0 is rejected if and only if $Y \geq b_{\alpha/2}$ or $Y \leq n - b_{\alpha/2} = b^l_{\alpha/2}$, the relevant NPI lower and upper probabilities, given $Y = y$, are also easily derived from (1.4) and (1.5) and are equal to

$$\begin{aligned}
&\underline{P}(Y_f \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}|y) = \\
&1 - \binom{2n}{n}^{-1} \times \left[\binom{2n-y}{n-y} + \sum_{l=1}^{b^l_{\alpha/2}} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right. \\
&+ \left\{ \binom{y+b_{\alpha/2}}{y} - \binom{y+b^l_{\alpha/2}}{y} \right\} \times \binom{2n-y-b_{\alpha/2}}{n-y} \\
&\left. + \sum_{l=b_{\alpha/2}+1}^{n} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right] \\
&\overline{P}(Y_f \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}|y) = \\
&\binom{2n}{n}^{-1} \times \left[\binom{y+b^l_{\alpha/2}+1}{y} \binom{2n-y-b^l_{\alpha/2}-1}{n-y} \right. \\
&\left. + \sum_{l=b^l_{\alpha/2}+2}^{b_{\alpha/2}-1} \left\{ \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right\} \right]
\end{aligned}$$

The NPI lower and upper reproducibility probabilities, in this case are as follows. For $y \leq b^l_{\alpha/2}$ or $y \geq b_{\alpha/2}$, so in case the original test led to rejection of H_0, we have $\underline{RP}(y) = \underline{P}(Y_f \leq b^l_{\alpha/2} \vee Y_f \geq b_{\alpha/2}|y) = 1 - \overline{P}(Y_f \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}|y)$ and $\overline{RP}(y) = \overline{P}(Y_f \leq b^l_{\alpha/2} \vee Y_f \geq b_{\alpha/2}|y) = 1 - \underline{P}(Y_f \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}|y)$. For $y \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}$, which led to H_0 not being rejected in the original test, we have $\underline{RP}(y) = \underline{P}(Y_f \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}|y)$ and $\overline{RP}(y) = \overline{P}(Y_f \in \{b^l_{\alpha/2}+1, \ldots, b_{\alpha/2}-1\}|y)$.

These results for NPI-RP for the one-sample sign test are illustrated and discussed in the Example 3.1.

Now we prove that the minimum value of the NPI lower reproducibility probability is equal to 0.5.

Proof of minimum value for lower RP

The NPI upper and lower probabilities for Bernoulli random quantities, as given in Equations (1.4) and (1.5), were derived by [16] through direct counting arguments. A nice alternative counting argument to derive these results is as follows (for more details we refer to [1, Section 2.2]). In the latent variable representation of Bernoulli data using real-valued outcomes of an experiment, with data consisting of n observations and interest in m future observations, the $\binom{n+m}{n}$ different orderings of these observations, when not distinguishing between the n observed values nor between the m future observations, are all equally likely. For each such ordering, the success-failure threshold can be in any of the $n+m+1$ intervals of the partition of the real line created by the $n+m$ values of the latent variables, leading to $n+m+1$ possible combinations (s, r), with s successes in the n tests and r successes in the m future observations. For such an ordering, these possible pairs (s, r) can be represented as a path on the rectangular lattice from $(0, 0)$ to (n, m) with steps going either one to the right or one upwards.

The $\binom{n+m}{n}$ different orderings, which are all equally likely, correspond to the $\binom{n+m}{n}$ different right-upwards paths from $(0, 0)$ to (n, m), and hence the NPI upper and lower probabilities (1.4) and (1.5) can also be derived by counting paths. To derive the NPI lower probability $\underline{P}(Y_{n+1}^{n+m} \in R_t|Y_1^n = s)$, one counts all such paths

which for given s must go only through points (s, r) with $r \in R_t$, so they do not go through (s, l) for any $l \in R_t^c$. The corresponding NPI upper probability $\overline{P}(Y_{n+1}^{n+m} \in R_t | Y_1^n = s)$ is derived by counting all such paths that go through at least one (s, r) with $r \in R_t$.

In Section 3.4 the following NPI lower probabilities are used, when considering $m = n$ future observations based on n data observations

$$\underline{P}(Y_{n+1}^{2n} \leq y | Y_1^n = y) = 0.5 \quad \text{for} \quad y \in \{0, 1, \ldots, n-1\} \tag{3.17}$$

$$\underline{P}(Y_{n+1}^{2n} \geq y | Y_1^n = y) = 0.5 \quad \text{for} \quad y \in \{1, 2, \ldots, n\} \tag{3.18}$$

These lower probabilities follow from symmetry arguments as follows. The lower probability (3.17) is derived by counting all paths which go through (y, r) with $r \leq y$ but do not go through $(y, y+1)$, which then implies that they also do not go through (y, t) for any $t \geq y+1$. To ensure that the paths are not counted more than once, one can count the paths going through (y, u) and $(y+1, u)$, for a specific $u \in \{0, 1, \ldots, y\}$, and sum all these paths for these values of u. By symmetry, these are precisely half of all the $\binom{2n}{n}$ different right-upwards paths from $(0, 0)$ to (n, n), which follows by considering the paths that go through a pair (v, y) and $(v, y+1)$ for any $v \in \{0, 1, \ldots, y\}$, these are precisely the paths that are not counted in deriving the lower probability (3.17). Every right-upwards path from $(0, 0)$ to (n, n) goes either through (y, u) and $(y+1, u)$, for a specific $u \in \{0, 1, \ldots, y\}$, or through (v, y) and $(v, y+1)$ for a specific $v \in \{0, 1, \ldots, y\}$. By symmetry it follows that there are exactly the same number of paths going through (y, u) and $(y+1, u)$ for a specific $u \in \{0, 1, \ldots, y\}$ as paths through (u, y) and $(u, y+1)$ for the same specific value $u \in \{0, 1, \ldots, y\}$ (this can be seen by replacing every upwards step by a right step and vice versa). This proves the result (3.17).

The derivation of the lower probability (3.18) is, interestingly, almost given by the previous counting argument, as we have seen that precisely half of all paths go through (v, y) and $(v, y+1)$ for a specific $v \in \{0, 1, \ldots, y\}$. To determine (3.18), we need to count all right-upwards paths from $(0, 0)$ to (n, n) that go through (y, r) with $r \geq y$ but not through $(y, y-1)$, which implies that they also do not go through (y, t) for any $t \leq y-1$. These paths are identical to all the paths which go through

(v, y) and $(v, y+1)$ for a specific $v \in \{0, 1, \ldots, y\}$, but with one exception, namely we have to exclude the paths that go through $(y, y-1)$ and (y, y) and $(y, y+1)$. However, we must include in this counting argument the paths that go through $(y-1, y)$ and (y, y) and $(y+1, y)$, and again by symmetry this means that we have to include the same number of paths as we just had to exclude. Hence, half of all such paths from $(0, 0)$ to (n, n) are included in the count to derive (3.18), which concludes the proof of this result.

Example 3.1

We consider NPI-RP for the one-sample sign test with either $n = 20$ or $n = 30$ observations and with either $\alpha = 0.05$ or $\alpha = 0.01$. First, we consider the one-sided test with $H_1 : \theta > 0$ (we do not illustrate cases with $H_1 : \theta < 0$, these follow by symmetry and therefore have similar behaviour), thereafter we consider the two-sided test with $H_1 : \theta \neq 0$.

Table 3.1 presents the NPI lower and upper reproducibility probabilities, $\underline{RP}(y)$ and $\overline{RP}(y)$, for the one-sided test with $H_1 : \theta > 0$ and with $n = 20$ observations in the first test, of which y were positive. In Table 3.1 the level of significance is $\alpha = 0.05$, which leads to the null-hypothesis being rejected if and only if $y \geq 15$. The results for the corresponding test but with $\alpha = 0.01$ are presented in Table 3.2, in this case the null-hypothesis is rejected if and only if $y \geq 16$. In all tables in this thesis the entries are rounded to three decimals, but fewer decimals are given if the values are exact. The NPI upper probability for RP is equal to 1 in case $y = 0$ or $y = 20$, as discussed previously this reflects that such data do not provide evidence against the possible situation that there would never be any positive values (for $y = 0$), which would certainly lead to reproducibility of non-rejection of the null-hypothesis, or that there would never be any negative values (for $y = 20$), which would certainly lead to reproducibility of rejection of the null-hypothesis. The NPI lower probabilities for RP are exactly equal to 0.5 for $y = 14$ and $y = 15$ in Table 3.1 and for $y = 15$ and $y = 16$ in Table 3.2, this was also discussed previously and is proven in Proof of minimum value for lower RP . Of course, for $\alpha = 0.01$ the NPI lower and upper reproducibility probabilities, for y such that the null-hypothesis is

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1	7	0.988	0.995	14	0.5	0.634
1	1.000	1.000	8	0.973	0.988	15	0.5	0.642
2	1.000	1.000	9	0.947	0.973	16	0.642	0.775
3	1.000	1.000	10	0.905	0.947	17	0.775	0.882
4	0.999	1.000	11	0.840	0.905	18	0.882	0.954
5	0.998	0.999	12	0.750	0.840	19	0.954	0.990
6	0.995	0.998	13	0.634	0.750	20	0.990	1

Table 3.1: NPI-RP for Sign test with $H_1 : \theta > 0$, $n = 20$, $\alpha = 0.05$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1	7	0.995	0.998	14	0.642	0.760
1	1.000	1.000	8	0.989	0.995	15	0.5	0.642
2	1.000	1.000	9	0.976	0.989	16	0.5	0.653
3	1.000	1.000	10	0.952	0.976	17	0.653	0.796
4	1.000	1.000	11	0.912	0.952	18	0.796	0.909
5	0.999	1.000	12	0.850	0.912	19	0.909	0.976
6	0.998	0.999	13	0.760	0.850	20	0.976	1

Table 3.2: NPI-RP for Sign test with $H_1 : \theta > 0$, $n = 20$, $\alpha = 0.01$

not rejected, are greater than for $\alpha = 0.05$, while for y such that the null-hypothesis is rejected the reverse holds; these properties follow directly from the fact that the null-hypothesis is rejected for fewer values for y if α is smaller. As discussed in Section 1.4, there has been some confusion between RP and significance levels. It is clear from these tables that they are very different concepts without a direct relation, while it is also clearly important to take the actual test data from the first test (here the value y) into account as the RP inferences depend strongly on the actual test data, which is fully in line with intuition.

Tables 3.3 and 3.4 present the NPI-RP results for the same situations as in Tables 3.1 and 3.2, but with $n = 30$ instead of $n = 20$. In both these tables, entries for small positive values of y with $\underline{RP}(y) = 1.000$ and $\overline{RP}(y) = 1.000$ are not included in the tables. For $\alpha = 0.05$ in Table 3.3, the null-hypothesis is rejected if and only if $y \geq 20$, while for $\alpha = 0.01$ in Table 3.4, the null-hypothesis is rejected if and only if $y \geq 22$. Of course, these tables illustrate the same aspects of the NPI-RP approach as discussed above, and the comparison with Tables 3.1 and 3.2 also shows that the imprecision, the difference between corresponding upper and lower probabilities, is somewhat smaller with $n = 30$ than with $n = 20$, which is in line with the general behaviour of statistical inferences in NPI [19].

Tables 3.5-3.7 present the NPI-RP results for the one-sample sign test with the

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1	15	0.853	0.904	23	0.801	0.874
7	0.999	1.000	16	0.785	0.853	24	0.874	0.928
8	0.998	0.999	17	0.702	0.785	25	0.928	0.964
9	0.995	0.998	18	0.605	0.702	26	0.964	0.985
10	0.990	0.995	19	0.5	0.605	27	0.985	0.995
11	0.981	0.990	20	0.5	0.608	28	0.995	0.999
12	0.965	0.981	21	0.608	0.710	29	0.999	1.000
13	0.941	0.965	22	0.710	0.801	30	1.000	1
14	0.904	0.941						

Table 3.3: NPI-RP for Sign test with $H_1 : \theta > 0$, $n = 30$, $\alpha = 0.05$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.000	1	16	0.910	0.945	24	0.724	0.820
9	0.999	1.000	17	0.861	0.910	25	0.820	0.895
10	0.998	0.999	18	0.794	0.861	26	0.895	0.948
11	0.996	0.998	19	0.710	0.794	27	0.948	0.979
12	0.991	0.996	20	0.611	0.710	28	0.979	0.994
13	0.982	0.991	21	0.5	0.611	29	0.994	0.999
14	0.968	0.982	22	0.5	0.614	30	0.999	1
15	0.945	0.968	23	0.614	0.724			

Table 3.4: NPI-RP for Sign test with $H_1 : \theta > 0$, $n = 30$, $\alpha = 0.01$

two-sided alternative hypothesis $H_1 : \theta \neq 0$, for cases that correspond to those in Tables 3.1-3.4, but with the case $n = 30$ only presented once, in Table 3.7, for $\alpha = 0.01$. In Table 3.7, the values are not included for $y \geq 17$, these are equal to the values for $30 - y$ in the same table as follows from symmetry (and as illustrated by the entries for $y = 16$ and $y = 14$).

In the case of Table 3.5, with $n = 20$ and $\alpha = 0.05$, the null-hypothesis is rejected if and only if $y \leq 5$ or $y \geq 15$. At these values the lower and upper reproducibility probabilities are minimal, which is of course logical, and it is important to notice that these lower probabilities are now not exactly equal to 0.5, this is simply because of the two rejection areas for the null-hypothesis, that is for both small and large values of y, hence the events of interest here are not similar to those discussed in Proof of minimum value for lower RP . Indeed, compared to those events in Proof of minimum value for lower RP (and those that led to lower probabilities 0.5 in Tables 3.1-3.4), the minimal value for $\underline{RP}(y)$ for y which leads to rejection of the null-hypothesis, namely 0.501 at $y = 5$ and $y = 15$, is greater than 0.5. This is due to the small but possible case that a future test might lead to rejection of the null-hypothesis for what could be considered as 'opposite reasons', namely where the

first test is rejected with small value of y but the future test with a large value of Y_f, and vice versa. This small possibility of getting future observations at the 'other end' leading to rejection of the null-hypothesis is also the reason for the minimal for $\underline{RP}(y)$ for y which leads to non-rejection of the null-hypothesis, namely 0.495 at $y = 6$ and $y = 14$, to be less than 0.5. In Sections 3.5 and 3.6 we will encounter more situations with the NPI lower reproducibility probability being less than 0.5, in some cases substantially so; this feature of the NPI approach to RP will be discussed in the examples in these sections.

Table 3.5 also clearly illustrates the symmetry in these inferences about $y = n/2 = 10$. Notice further that these NPI lower and upper reproducibility probabilities for values of y for which the null-hypothesis is not rejected, are maximally equal to 0.809 and 0.895, respectively, so quite substantially less than $1 - \alpha$, which again emphasizes that the RP and significance level are different concepts which one must be careful not to interpret wrongly. The NPI upper probabilities in Tables 3.5-3.7 are exactly equal to 1 for $y = 0$ and $y = n$, which here both imply rejection of the null-hypothesis and hence that such data do not provide evidence against the possibility that the observations will always be either all negative or all positive.

For the case in Table 3.6, with $n = 20$ and $\alpha = 0.01$, the null-hypothesis is rejected if and only if $y \leq 3$ or $y \geq 17$, and compared to Table 3.5, the NPI lower and upper reproducibility probabilities are now again larger for values of y for which the null-hypothesis is not rejected and smaller for values of y for which the null-hypothesis is rejected. This is logical and in line with the same feature as discussed for the one-sided tests above. Finally, for the case in Table 3.7, with $n = 30$ and $\alpha = 0.01$, the null-hypothesis is rejected if and only if $y \leq 7$ or $y \geq 23$, and the entries in this table, when compared to those in Table 3.6, show again less imprecision as a result of having more data observations, as was also discussed for the one-sided tests.

The NPI-RP approach is strongly based on the test data, and as such reproducibility tends to increase with larger sample size, reflected in larger values of the NPI lower and upper reproducibility probabilities. The comparable cases with $n = 20$ and $n = 30$ in this example illustrate this. However, direct comparison is

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.990	1	7	0.622	0.745	14	0.495	0.633
1	0.954	0.990	8	0.723	0.827	15	0.501	0.644
2	0.882	0.954	9	0.787	0.878	16	0.642	0.775
3	0.775	0.883	10	0.809	0.895	17	0.775	0.883
4	0.642	0.775	11	0.787	0.878	18	0.882	0.954
5	0.501	0.644	12	0.723	0.827	19	0.954	0.990
6	0.495	0.633	13	0.622	0.745	20	0.990	1

Table 3.5: NPI-RP for Sign test with $H_1 : \theta \neq 0$, $n = 20$, $\alpha = 0.05$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.947	1	7	0.862	0.922	14	0.774	0.863
1	0.829	0.947	8	0.918	0.957	15	0.652	0.775
2	0.669	0.829	9	0.949	0.976	16	0.500	0.653
3	0.500	0.669	10	0.959	0.981	17	0.500	0.669
4	0.500	0.653	11	0.949	0.976	18	0.669	0.829
5	0.652	0.775	12	0.918	0.957	19	0.829	0.947
6	0.774	0.863	13	0.862	0.922	20	0.947	1

Table 3.6: NPI-RP for Sign test with $H_1 : \theta \neq 0$, $n = 20$, $\alpha = 0.01$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.998	1	6	0.619	0.734	12	0.862	0.913
1	0.987	0.998	7	0.500	0.620	13	0.906	0.944
2	0.960	0.987	8	0.500	0.614	14	0.932	0.962
3	0.910	0.960	9	0.614	0.716	15	0.940	0.967
4	0.833	0.910	10	0.715	0.800	16	0.932	0.962
5	0.734	0.833	11	0.800	0.866			

Table 3.7: NPI-RP for Sign test with $H_1 : \theta \neq 0$, $n = 30$ and $\alpha = 0.01$.

slightly difficult because with increasing n one has, of course, a larger range of y values to consider. One might compare similar ratios y/n, but care must be taken when close to the threshold for rejection of the null-hypothesis as the test results for similar ratios y/n, but with different values of n, might be different.

3.5 NPI-RP for the One Sample Signed Rank Test

The NPI-RP approach for this one-sided signed-rank test is as follows. Given n ordered data observations $x_{(1)} < ... < x_{(n)}$, we use NPI for n future observations $X_{n+1}, ..., X_{2n}$. As described in the previous sections and presented in detail by [5], there are $\binom{2n}{n}$ possible orderings of these n future observations among the n data observations, and all these orderings are equally likely to occur. The idea of these orderings is explained in Example 2.1 in Chapter 2. For each of these orderings we are interested in the test statistic W^f of the signed-rank test for the n future observations. As these future observations are not precise but only their number in each of the intervals of the partition created by the n data observations (from the first test) are known for a given ordering, we cannot calculate a single precise value of W^f related to an ordering, but we can deduce the minimum and maximum possible values; we denote these by $\underline{W}^f$ and $\overline{W}^f$, respectively. Doing this for each of the $\binom{2n}{n}$ different orderings then leads to NPI lower and upper reproducibility probabilities for this test.

For reproducibility in case the original experiment led to rejection of the null-hypothesis $H_0 : \theta = 0$, the test statistic W^f has to be greater than or equal to the critical value W_α. To find the NPI lower reproducibility probability in this case, we count all of the $\binom{2n}{n}$ orderings for which $W^f \geq W_\alpha$ must certainly hold, so for which $\underline{W}^f \geq W_\alpha$. In this same case, the corresponding NPI upper reproducibility probability is found by counting the orderings for which $W^f \geq W_\alpha$ can possibly hold, so for which $\overline{W}^f \geq W_\alpha$. If the first experiment does not lead to rejection of H_0, then the test result is reproduced if $W^f < W_\alpha$ and the corresponding NPI lower and upper reproducibility probabilities are derived as above but with the obvious change of inequality signs.

We next explain how the minimum and maximum values of W^f, $\underline{W}^f$ and $\overline{W}^f$, for a specific ordering can be calculated, detailed justification of this is presented in Proof of equation 3.19. The original ordered test data $x_{(1)} < ... < x_{(n)}$, together with definitions $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$ define $n+1$ intervals $(x_{(j-1)}, x_{(j)})$ for $j = 1, \ldots, n+1$. A specific ordering of these n data observations together with n future observations is specified by the numbers of these future observations in each of these intervals. Let S_j be the number of the n future observations in interval $(x_{(j-1)}, x_{(j)})$ for $j = 1, \ldots, n+1$. Of course, $S_j \geq 0$ and $\sum_{j=1}^{n+1} S_j = n$, and the assumptions $A_{(n)}, \ldots, A_{(2n-1)}$ underlying NPI for n future real-valued data [5] imply that all the different combinations of such values S_j, for $j = 1, \ldots, n+1$, are equally likely. We consider now one specific ordering, which is therefore specified by the values $(S_1, \ldots, S_{n+1})$.

The minimum possible value $\underline{W}^f$ of the test statistic for testing $H_0 : \theta = 0$, for n future observations corresponding to this specific ordering $(S_1, \ldots, S_{n+1})$, is easily seen to be achieved if all future observations are put at the left end-points of their respective intervals. As these are open intervals, it would be just a small positive value to the right of these left end-points, also to avoid ties we would seperate multiple values at the same point with a small positive value; while this is required for mathematical correctness, it would complicate the presentation while not affecting the results, so we will just use the expression 'at $x_{(j-1)}$' for all future observations in $(x_{(j-1)}, x_{(j)})$ that we wish to locate as far to the left as possible; and similar for the case where we want to put future observations as far to the right as possible when calculating the corresponding NPI-RP upper probability. Hence, for this specific ordering, S_j future observations are put at $x_{(j-1)}$ in order to get the minimum possible value $\underline{W}^f$ for the test statistic.

To proceed, we introduce further notation in order to distinguish between positive and negative values for the future observations which are put at these left end-points of the intervals. First, the absolute values of $x_{(0)} < x_{(1)} < \ldots < x_{(n)}$ are ordered, with ranks $j = 1, \ldots, n+1$, and we introduce the notation $x_{|j|}$ for this j-th ordered value if it is positive and $x_{-|j|}$ if it is negative (we neglect the possibility of an original observation being exactly equal to 0 as this is of little practical interest and would

complicate the presentation). Note that $x_{-|n+1|} = -\infty$. For $j = 1, \ldots, n+1$, let T_j be the number of future observations, in the specific ordering considered, that are put at $x_{|j|}$, and T_{-j} the number of such future observations that are put at $x_{-|j|}$. This means that $T_j = S_l$ with $x_{(l-1)} = x_{|j|} > 0$ and $T_{-j} = S_l$ with $x_{(l-1)} = x_{-|j|} < 0$. This leads to the minimum possible value for the test statistic W^f, that is possible corresponding to a specific ordering $(S_1, \ldots, S_{n+1})$, being equal to

$$\underline{W}^f = \sum_{j>0} T_j \left[\frac{T_j + 1}{2} + \sum_{|i|<j} T_i \right] \tag{3.19}$$

where the range of values for j in the summations is restricted to $1, \ldots, n$, and this is the range of i, but i can be negative or positive and can be from $-(n-1)$ to $(n-1)$ in order to have its absolute value less than j with j ranging from 1 to n. The justification of Equation (3.19) is given in the next Proof. It should be emphasized that this value $\underline{W}^f$ corresponds to a single specific ordering $(S_1, \ldots, S_{n+1})$, and that in total $\binom{2n}{n}$ such orderings have to be considered. If these $\binom{2n}{n}$ values $\underline{W}^f$ have been calculated, the proportion of these that are greater than or equal to W_α is equal to the NPI lower reproducibility probability in this case where the original test led to rejection of H_0.

The above results can be quite easily implemented in an algorithm, as we have done in order to illustrate this NPI-RP method in Example 3.2, but if n is not small the total computational effort soon becomes very substantial. Of course, one does not need to use Equation (3.19), but then one would need to order the future observations, still assuming them to be 'at' the left end-points of the intervals, and calculate the sum of signed-ranks, so using (3.19) simplifies this as it leaves out the need to order each one of the $\binom{2n}{n}$ future samples involved. There are several possible ways to deal with this, for example one can make the algorithm more efficient by using logical monotonicity relationships between the orderings and the corresponding values $\underline{W}^f$, this would lead to some combinatorial challenges but seems feasible. We will not pursue this further in this thesis, as the intention is to introduce NPI-RP and we restrict to illustrations involving small numbers of data. In Chapter 4 we are developing an alternative approach for such reproducibility inferences for the signed-rank test (and other tests), which is based on bootstrapping from a NPI

perspective.

To calculate the NPI upper reproducibility probability for the signed-rank test in this scenario, with the null-hypothesis rejected on the basis of the data observed in the first test, we follow the same steps as above except now all S_j future observations in the interval $(x_{(j-1)}, x_{(j)})$ are put at $x_{(j)}$, for $j = 1, \ldots, n+1$, leading to the maximum possible value $\overline{W}^f$ for the test statistic corresponding to the specific ordering $(S_1, \ldots, S_{n+1})$. This value is derived similarly to the derivation of $\underline{W}^f$ presented above, with just the changes due to the use of the right end-points instead of the left end-points of the intervals. Now, the absolute values of $x_{(1)} < x_{(2)} < \ldots < x_{(n+1)}$ are ordered, with ranks $j = 1, \ldots, n+1$, and we introduce the notation $\tilde{x}_{|j|}$ for this j-th ordered value if it is positive and $\tilde{x}_{-|j|}$ if it is negative. Note that $\tilde{x}_{|n+1|} = \infty$. For $j = 1, \ldots, n+1$, let $\tilde{T}_j$ be the number of future observations, in the specific ordering considered, that are put at $\tilde{x}_{|j|}$, and $\tilde{T}_{-j}$ the number of such future observations that are put at $\tilde{x}_{-|j|}$. This means that $\tilde{T}_j = S_l$ with $x_{(l)} = \tilde{x}_{|j|} > 0$ and $\tilde{T}_{-j} = S_l$ with $x_{(l)} = \tilde{x}_{-|j|} < 0$. This leads to the maximum possible value for the test statistic W^f, that is possible corresponding to a specific ordering $(S_1, \ldots, S_{n+1})$, being equal to

$$\overline{W}^f = \sum_{j>0} \tilde{T}_j \left[\frac{\tilde{T}_j + 1}{2} + \sum_{|i|<j} \tilde{T}_i \right] \tag{3.20}$$

where the range of values for j in the summations is restricted to $1, \ldots, n+1$, and this is the range of i, but i can be negative or positive and can be from $-(n-1)$ to $(n-1)$ in order to have its absolute value less than j with j ranging from 1 to n. It is clear that, apart from the differences in the definitions of the values $\tilde{T}_j$ and T_j, (3.20) is equal to (3.19); its derivation follows exactly the same steps as that of (3.19), which is given in Proof of equation 3.19.

The NPI lower and upper reproducibility probabilities for this test, in case the original data did not lead to rejection of H_0, can be derived similarly. It is easier to just calculate the NPI lower and upper probabilities for the event $W^f \geq W_\alpha$ as presented above, and then to use the conjugacy property (1.5) to derive the NPI lower and upper probabilities for the event $W^f < W_\alpha$, which corresponds to reproducing the original test result in this case.

While an analytic investigation of properties of this NPI-RP approach is very

difficult due to the absence of closed-form expressions for the lower and upper probabilities, one property is easy to derive: if α is very small, such that H_0 is only rejected if all observations are positive, and if the original test data indeed led to such rejection, then the NPI lower reproducibility probability is equal to 0.5 and the corresponding upper reproducibility probability is equal to 1, for all sample sizes n. This is based on the fact that exactly half of the possible orderings have one or more future observations smaller than X_1. Calculations of the lower NPI reproducibility probability assume such values to be negative. Example 3.2 illustrates the NPI-RP method for the signed-rank test as presented in this section.

Proof of Equation 3.19

We provide a justification for the minimum value of the test statistic for future data in case of the Wilcoxon one-sample rank-sum test, as given by Equation (3.19) in Section 3.5. The justification for the corresponding maximum value (3.20) follows precisely the same steps.

With the defined notation, for a specific ordering $(S_1, \ldots, S_{n+1})$ there are T_j future observations at $x_{|j|}$ and T_{-j} future observations at $x_{-|j|}$. We must sum the ranks of the positive observations, so the ranks of the T_j future observations at $x_{|j|}$ for the values $j \in \{1, \ldots, n+1\}$ which are such that they appear as index $|j|$ in these $x_{|j|}$. The ranks of the T_j future observations at $x_{|j|}$ are equal to

$$\sum_{|i|<j} T_i + 1, \quad \ldots, \quad \sum_{|i|<j} T_i + T_j$$

which sum up to

$$T_j \times \sum_{|i|<j} T_i + \sum_{k=1}^{T_j} k = T_j \times \sum_{|i|<j} T_i + \frac{T_j(T_j+1)}{2}$$

Summing these values for all T_j, so corresponding to all the intervals between consecutive original data observations with positive left end-points, gives the minimum value of the rank sum corresponding to this specific ordering, hence indeed

$$\underline{W}^f = \sum_{j>0} \frac{T_j(T_j+1)}{2} + \sum_{j>0} \sum_{|i|<j} T_j T_i \tag{3.21}$$

which leads to Equation (3.19).

Example 3.2

We illustrate the reproducibility of the one-sample signed-rank test with $H_0 : \theta = 0$ and $H_1 : \theta > 0$. We consider a variety of samples, each of size $n = 6$, and level of significance $\alpha = 0.05$. For this test, the null-hypothesis is rejected if $W \geq 19$. In the NPI approach, there are $\binom{12}{6} = 924$ orderings of 6 future observations among the 6 data observations to consider, and all are assumed to be equally likely by the assumptions underlying the method presented in this section. Table 3.8 presents the NPI lower and upper reproducibility probabilities for a variety of data represented by signed ranks. These cover many cases that can occur in practice, where of course aspects of the actual data values beyond their signed ranks are irrelevant. The first three cases represent the only possible data sets for which H_0 is rejected, for all the other cases H_0 is not rejected, which of course affects how reproducibility is defined.

If all six data observations are positive then $\underline{RP} = 0.5$ and $\overline{RP} = 1$. The latter is intuitively logical, as these data do not provide strong evidence against the possibility that the data in the process considered will never be negative. This NPI lower probability being equal to 0.5 results from the fact that, for any ordering with (at least) one observation smaller than the smallest data observation, that future observation is not restricted in its value, in the sense that for the configuration related to this lower probability we effectively put such an observation at $-\infty$, hence it will have signed rank -6. This means that in this case, for each ordering with one or more future observations smaller than the smallest data observation, the minimal signed-rank sum for the future observations will be at most 15. Half of the 924 orderings have at least one such smaller future observation, which is in line with intuition, because with the data set and a future sample of the same size n, the probability for the event that the future sample contains the smallest of all $2n$ values, is equal to 0.5. It is further interesting to note that $\underline{RP}$ is substantially less than 0.5 for several of the possible cases. This implies that such data, typically with test statistics close to the critical value $W_{0.05} = 19$, do not provide strong evidence in favour of the reproducibility of the test result. Of course, this is based on the fact that the lower RP corresponds to the situation where, for each of the 924 possible orderings, positive future values are assumed to be as small as possible within the

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5,6	21	0.5	1
-1,2,3,4,5,6	20	0.364	0.773
-2,1,3,4,5,6	19	0.326	0.712
-3,1,2,4,5,6	18	0.364	0.718
-2,-1,3,4,5,6	18	0.5	0.788
-4,1,2,3,5,6	17	0.429	0.750
-3,-1,2,4,5,6	17	0.538	0.810
-5,1,2,3,4,6	16	0.472	0.767
-3,-2,1,4,5,6	16	0.576	0.831
-4,-1,2,3,5,6	16	0.581	0.831
-6,1,2,3,4,5	15	0.494	0.773
-3,-2,-1,4,5,6	15	0.728	0.902
-6,-3,-1,2,4,5	11	0.805	0.935
-4,-3,-2,-1,5,6	11	0.879	0.966
-6,-5,-4,-3,1,2	3	0.957	0.992
-6,-5,-4,-2,-1,3	3	0.973	0.997
-6,-5,-4,-3,-2,-1	0	0.992	1

Table 3.8: NPI-RP for signed-rank test with $H_1 : \theta > 0$, $n = 6$, $\alpha = 0.05$, $W_{0.05} = 19$.

interval to which they belong according to the specific ordering, and this also holds for the negative future values, hence the negative values tend to get larger absolute ranks.

The entries in Table 3.8 show another interesting feature, namely that $\underline{RP}$ and $\overline{RP}$ are not monotone as functions of the difference between the actual data test statistic W and W_α. For each specific value $W < 19$, $\underline{RP}$ and $\overline{RP}$ are minimal for the actual data with the smallest number of negative values. If one restricts attention to data sets with the same number of negative signed ranks, then the monotonicity holds. So the NPI-RP method takes both the actual signed ranks and the number of negative values into account, which seems a particularly nice feature of the method.

We have performed the same calculations for more orderings with $n = 6$, and similarly for $n = 4, 5, 7$. For the latter case, there are already $\binom{14}{7} = 3432$ orderings. The NPI-RP results for all these cases were similar as presented and discussed here for $n = 6$. Slightly larger values of n are also possible using our method, but as we calculate $\underline{W}^f$ and $\overline{W}^f$ for each ordering, and e.g. for $n = 10$ there are $\binom{20}{10} = 184,756$ orderings, the computational effort soon becomes too large. Nevertheless, the RP approach presented here shows some interesting properties and is, in principle, applicable for all data sets.

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5,6	21	0.5	1
-1,2,3,4,5,6	20	0.5	0.773
-2,1,3,4,5,6	19	0.5	0.773
-3,1,2,4,5,6	18	0.5	0.773
-2,-1,3,4,5,6	18	0.773	0.909
-4,1,2,3,5,6	17	0.5	0.773
-3,-1,2,4,5,6	17	0.773	0.909
-5,1,2,3,4,6	16	0.5	0.773
-3,-2,1,4,5,6	16	0.773	0.909
-4,-1,2,3,5,6	16	0.773	0.909
-6,1,2,3,4,5	15	0.5	0.773
-3,-2,-1,4,5,6	15	0.910	0.970
-6,-3,-1,2,4,5	11	0.910	0.970
-4,-3,-2,-1,5,6	11	0.970	0.992
-6,-5,-4,-3,1,2	3	0.970	0.992
-6,-5,-4,-2,-1,3	3	0.992	0.999
-6,-5,-4,-3,-2,-1	0	0.999	1

Table 3.9: NPI-RP for signed-rank test with $H_1 : \theta > 0$, $n = 6$, $\alpha = 0.016$, $W_{0.016} = 21$.

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4	10	0.786	1
-1,2,3,4	9	0.586	0.929
-2,1,3,4	8	0.543	0.871
-3,1,2,4	7	0.514	0.829
-2,-1,3,4	7	0.371	0.757
-4,1,2,3	6	0.5	0.800
-3,-1,2,4	6	0.329	0.700
-4,-1,2,3	5	0.357	0.714
-3,-2,1,4	5	0.371	0.714
-4,-2,1,3	4	0.414	0.743
-3,-2,-1,4	4	0.5	0.843
-4,-3,1,2	3	0.457	0.757
-4,-2,-1,3	3	0.557	0.886
-4,-3,-1,2	2	0.614	0.914
-4,-3,-2,1	1	0.671	0.929
-4,-3,-2,-1	0	0.786	1

Table 3.10: NPI-RP for signed-rank test with $H_1 : \theta > 0$, $n = 4$, $\alpha = 0.438$, $W_{0.438} = 6$.

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4	10	0.5	1
-1,2,3,4	9	0.5	0.786
-2,1,3,4	8	0.5	0.786
-3,1,2,4	7	0.5	0.786
-2,-1,3,4	7	0.786	0.929
-4,1,2,3	6	0.5	0.786
-3,-1,2,4	6	0.786	0.929
-4,-1,2,3	5	0.786	0.929
-3,-2,1,4	5	0.786	0.929
-4,-2,1,3	4	0.786	0.929
-3,-2,-1,4	4	0.929	0.986
-4,-3,1,2	3	0.786	0.929
-4,-2,-1,3	3	0.929	0.986
-4,-3,-1,2	2	0.929	0.986
-4,-3,-2,1	1	0.929	0.986
-4,-3,-2,-1	0	0.986	1

Table 3.11: NPI-RP for signed-rank test with $H_1 : \theta > 0$, $n = 4$, $\alpha = 0.062$, $W_{0.062} = 10$.

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5,6,7	28	0.5	1
-1,2,3,4,5,6,7	27	0.439	0.904
-2,1,3,4,5,6,7	26	0.396	0.834
-3,1,2,4,5,6,7	25	0.358	0.773
-4,1,2,3,5,6,7	24	0.319	0.709
-3,-1,2,4,5,6,7	24	0.294	0.661
-5,1,2,3,4,6,7	23	0.359	0.721
-3,-2,1,4,5,6,7	23	0.402	0.741
-6,1,2,3,4,5,7	22	0.425	0.753
-3,-2,-1,4,5,6,7	22	0.5	0.794
-7,1,2,3,4,5,6	21	0.475	0.769
-4,-3,1,2,5,6,7	21	0.498	0.795
-4,-3,-2,-1,5,6,7	18	0.704	0.896
-7,-6,1,2,3,4,6	15	0.736	0.904
-5,-4,-3,-2,-1,6,7	13	0.857	0.961
-7,-6,-5,1,2,3,4	10	0.872	0.965
-6,-5,-4,-3,-2,-1,7	7	0.949	0.992
-7,-6,-5,-4,1,2,3	6	0.939	0.990
-7,-6,-5,-4,-3,-2,-1	0	0.990	1

Table 3.12: NPI-RP for signed-rank test with $H_1 : \theta > 0$, $n = 7$, $\alpha = 0.055$, $W_{0.055} = 24$.

sign-ranked data	W	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5,6,7	28	0.5	1
-1,2,3,4,5,6,7	27	0.5	0.769
-2,-1,3,4,5,6,7	26	0.769	0.904
-3,1,2,4,5,6,7	25	0.5	0.769
-4,1,2,3,5,6,7	24	0.5	0.769
-3,-1,2,4,5,6,7	24	0.769	0.904
-5,1,2,3,4,6,7	23	0.5	0.769
-3,-2,1,4,5,6,7	23	0.769	0.904
-6,1,2,3,4,5,7	22	0.5	0.769
-3,-2,-1,4,5,6,7	22	0.904	0.965
-7,1,2,3,4,5,6	21	0.5	0.769
-4,-3,1,2,5,6,7	21	0.769	0.904
-4,-3,-2,-1,5,6,7	18	0.965	0.990
-7,-6,1,2,3,4,6	15	0.769	0.904
-5,-4,-3,-2,-1,6,7	13	0.990	0.998
-7,-6,-5,1,2,3,4	10	0.904	0.965
-6,-5,-4,-3,-2,-1,7	7	0.998	0.999
-7,-6,-5,-4,1,2,3	6	0.965	0.990
-7,-6,-5,-4,-3,-2,-1	0	0.999	1

Table 3.13: NPI-RP for signed-rank test with $H_1 : \theta > 0$, $n = 7$, $\alpha = 0.008$, $W_{0.008} = 28$.

The results are similar for $n = 6$ with $\alpha = 0.016$, as shown in Table 3.9, in this case H_0 is rejected if $W \geq 21$. In Table 3.9 we note that the values of $\overline{RP}$ and $\underline{RP}$ are similar in some cases, for example in case 2, 3, 4, 6, 8, 11 the values $\underline{RP} = 0.5$ and $\overline{RP} = 0.773$ when we take a look at future values we can see that the test statistic W of these cases are different but the number of negative values are similar. With $n = 4$ and level of significance $\alpha = 0.438$. For this case, there are $\binom{8}{4} = 70$ orderings of 4 future observations among the 4 data observations, and H_0 is rejected if $W \geq 6$. Table 3.10 shows the NPI lower and upper reproducibility probabilities for many cases, the null hypothesis is rejected only in the first seven cases. If all 4 data observations are positive, the $\overline{RP} = 1$ but there is a different situation with $\underline{RP}$ which is larger than 0.5, it is 0.786 and this appears if the difference between the test statistic and the critical value is big (or α is large), in other ways, there are 79% of the 70 orderings have at least one observation smaller than the smallest data values. As we discussed earlier, the monotonicity of bounds holds with the same number of negative signed ranks, so the test statistic and the number of negative values are important in this method. When $\alpha = 0.062$ with $n = 4$ is used, the results are shown in Table 3.11, the properties of monotonicity and similarity of lower and upper values of RP which were discussed before show here. H_0 is reject if $W \geq 10$.

For $n = 7$, there are already $\binom{14}{7} = 3432$ orderings. The NPI-RP results for all these cases were similar as presented and discussed here for $n = 6$ and $n = 4$. Table 3.12 shows some of these cases with $\alpha = 0.055$, H_0 is rejected if and only if $W \geq 24$. Here, there are some cases that have $\underline{RP}$ less than 0.5 especially in cases that have W close to the critical value. For $\alpha = 0.008$, H_0 is rejected if $W \geq 28$ and the results are shown in Table 3.13, these results illustrate the same pattern of bounds of RP.

NPI-RP can also be developed for the two-sided signed-rank test, with $H_0 : \theta = \theta_0$ and alternative hypothesis $H_1 : \theta \neq \theta_0$, in which case the null-hypothesis is rejected if the test statistic is either in the $\alpha/2$ left or right tail of the null distribution. However, in this case it will not be easy to derive the minimum and maximum values for W^f, for a specific ordering, as it is not clear which configurations lead to these extremes; so it is not clear whether to put the S_j future observations within interval $(x_{(j-1)}, x_{(j)})$ all to one of the end-points or not. While some further theoretical results might be achievable, we think that the NPI-RP approach using bootstrapping, as we are developing in the next chapter, is more promising for such two-sided tests. In addition, in most practical scenarios, the real interest when using such a test is in providing an argument in favour of a specific one-sided alternative hypothesis, e.g. related to a new medication performing better than an established one, so development of the NPI-RP approach for the two-sided signed-rank test is likely to be of less practical value than for the single-sided test.

3.6 NPI-RP for the Two Sample Rank Sum Test

The NPI approach to reproducibility of this one-sided upper-tail two-sample rank sum test can be described as follows. The actual test with ordered data $x_{(1)} < \ldots < x_{(m)}$ and $y_{(1)} < \ldots < y_{(n)}$ leads to the test statistic Z, and to rejection of H_0 if $Z \geq Z_\alpha$ and non-rejection of H_0 if $Z < Z_\alpha$. As described in Section 1.4, we can use NPI for m future observations $X_{m+1}, \ldots, X_{2m}$ based on the information from data $x_{(1)} < \ldots < x_{(m)}$, where all $\binom{2m}{m}$ possible orderings of the m future X-observations and the m real X-observations are equally likely. Similarly, we can

use NPI for n future observations $Y_{n+1}, \ldots, Y_{2n}$ based on the information from data $y_{(1)} < \ldots < y_{(n)}$, with all $\binom{2n}{n}$ possible orderings equally likely. We now consider all $\binom{2m}{m}\binom{2n}{n}$ possible combinations of these different orderings, which again are all equally likely, based on the assumptions underlying NPI. For each combination, without further assumptions about the exact location of future observations within the intervals between consecutive real observations, the corresponding rank sum test statistic, denoted by Z^f, can take on one or more values. Suppose that the original data led to rejection of H_0, then this test result is certainly reproduced according to each such combination for which Z^f must be larger than or equal to Z_α, while it could possibly be reproduced according to all combinations for which Z^f can be greater than or equal to Z_α. So, in this case the NPI lower reproducibility probability is derived by counting the combinations for which $Z^f \geq Z_\alpha$ must certainly hold, while the corresponding NPI upper reproducibility probability is derived by counting the combinations for which this event can hold. Of course, if the original test leads to non-rejection of H_0, so if $Z < Z_\alpha$, then these NPI lower and upper reproducibility probabilities are derived by similar counting arguments but for the event $Z^f < Z_\alpha$.

As for the one-sample signed-rank test, presented in Section 3.5, we can derive the minimum and maximum values of Z^f corresponding to a specific ordering of the X and Y observations from the real test. This can be used to reduce the total computational effort, which however remains very substantial due to the very large number of combinations of such orderings that must be considered. We denote these minimum and maximum values of Z^f by $\underline{Z^f}$ and $\overline{Z^f}$, respectively. The minimum value $\underline{Z}^f$ is derived as follows. Let a specific ordering of the m future X observations among the corresponding m data observations be denoted by $(S_1^X, \ldots, S_{m+1}^X)$, where notation is in line with that introduced in Section 3.5, and let a specific ordering of the n future Y observations among the corresponding n data observations be denoted by $(S_1^Y, \ldots, S_{n+1}^Y)$. Furthermore, let $j(l) = \max\{j : x_{(j)} < y_{(l)}\}$ for $l = 1, \ldots, n+1$ and $j = 0, 1, \ldots, m$, so $x_{(j(l))} < y_{(l)} < x_{(j(l)+1)}$ and the rank of $y_{(l)}$ in the combined ordered data from both groups X and Y is $l + j(l)$. The minimum value for the

rank sum in this case is

$$\underline{Z}^f = \sum_{l=1}^{n+1} S_l^Y \left\{ \sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l-1)-1} S_t^X + \frac{S_l^Y + 1}{2} \right\} \tag{3.22}$$

the detailed justification for this result is presented in the next Proof. The corresponding maximum value is

$$\overline{Z}^f = \sum_{l=1}^{n+1} S_l^Y \left\{ \sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l)} S_t^X + \frac{S_l^Y + 1}{2} \right\} \tag{3.23}$$

which is also justified in the next Proof . To derive the NPI lower reproducibility probability for this scenario, all of the $\binom{2m}{m}\binom{2n}{n}$ equally likely possible orderings for which $\underline{Z}^f \geq Z_\alpha$ holds are counted and the total number is divided by $\binom{2m}{m}\binom{2n}{n}$, while the corresponding NPI upper reproducibility probability is derived by counting all of these possible orderings for which $\overline{Z}^f \geq Z_\alpha$ holds, and also dividing the total number by $\binom{2m}{m}\binom{2n}{n}$.

The NPI-RP approach for this test for the other scenarios of interest, with the initial test not rejecting the null-hypothesis or different one-sided alternative hypothesis, can be similarly derived, where NPI lower and upper reproducibility probabilities in case the original data did not lead to rejection of the null-hypothesis can, as in Section 3.5, be calculated using the results above and the conjugacy property (1.5). For the two-sided test the situation is more complex as the configurations leading to the minimum and maximum possible values of the test statistic, for given orderings, are not easy to derive anymore; for this case we consider the bootstrap based NPI approach, which we are developing and presenting in Chapter 4, more promising, particularly also as it can be applied with any sample sizes without the computational complexity of the method presented here. However, we do consider the NPI-RP method presented in this chapter to be important because the results are exact (the bootstrap-based method will only give approximate results and will require further assumptions), and while computation can be very substantial for the rank-based tests in this and the previous sections, for a given test result one only has to perform the computation once and it is easy to implement, so the method can be applied if one wishes to do so. As mentioned before, it will be possible to substantially reduce the computational effort by counting the orderings for which,

for example, $\underline{Z}^f \geq Z_\alpha$ holds, differently, using monotonicity properties of $\underline{Z}^f$ as function of the orderings $(S_1^X, \ldots, S_{m+1}^X)$ and $(S_1^Y, \ldots, S_{n+1}^Y)$. We have not pursued this further but it is an interesting topic for future research.

As for the test in Section 3.5, properties of the NPI-RP method for the two-sample rank sum test are difficult to investigate analytically. However, if α is so small that H_0 is only rejected if all n Y observations are greater than all m X observations, then $\underline{RP} = 0.25$ and $\overline{RP} = 1$. These values can be explained similarly as for the corresponding special case in Section 3.5, where here the lower probability corresponds to the case where all m future X observations are smaller than the largest observed X observation, and all n future Y observations are greater than the smallest observed Y observation. Both these events have probabilities 0.5 in the NPI framework, with the independence between the two groups leading to lower probability $0.5 \times 0.5 = 0.25$. This NPI-RP method for the two-sample rank sum test is illustrated and discussed in Example 3.3.

Proof of Equations 3.22 and 3.23

The derivation of the minimum value $\underline{Z}^f$ for the one-sided two-sample rank sum test in Section 3.6 is as follows, using the notation introduced in that section. To derive this minimum value, all S_j^X future X observations in the interval $(x_{(j-1)}, x_{(j)})$ are put at $x_{(j)}$ and all S_l^Y future Y observations in the interval $(y_{(l-1)}, y_{(l)})$ are put at $y_{(l-1)}$. The ranks of the S_l^Y future Y observations that are put at $Y_{(l-1)}$, in this configuration, are ranging from $\sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l-1)-1} S_t^X + 1$ to $\sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l-1)-1} S_t^X + S_l^Y$, so the contribution to the minimum total rank sum of these S_l^Y future observations is

$$S_l^Y \times \left(\sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l-1)-1} S_t^X \right) + \frac{S_l^Y (S_l^Y + 1)}{2}$$

The minimum total rank sum $\underline{Z}^f$, as given in Equation (3.22), follows by summing this for all $l = 1, \ldots, n+1$.

The corresponding maximum value $\overline{Z}^f$ is similarly derived, with all S_j^X future X observations in the interval $(x_{(j-1)}, x_{(j)})$ put at $x_{(j-1)}$ and all S_l^Y future Y observations in the interval $(y_{(l-1)}, y_{(l)})$ put at $y_{(l)}$. The ranks of the S_l^Y future Y observations

that are put at $Y_{(l)}$, in this configuration, are ranging from $\sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l)} S_t^X + 1$ to $\sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l)} S_t^X + S_l^Y$, so the contribution to the maximum total rank sum of these S_l^Y future observations is

$$S_l^Y \times \left(\sum_{k=1}^{l-1} S_k^Y + \sum_{t=1}^{j(l)} S_t^X \right) + \frac{S_l^Y (S_l^Y + 1)}{2}$$

and the maximum total rank sum $\overline{Z}^f$, as given in Equation (3.23), follows by summing this for all $l = 1, \ldots, n+1$.

Example 3.3

We illustrate NPI-RP for the two-sample rank sum test with $H_1 : \delta > 0$, with $m = n = 5$ data observations for both groups X and Y in several different rank orderings, and with $\alpha = 0.05$ which corresponds to critical value of the test statistic $Z_{0.05} = 36$, so H_0 is rejected if $Z \geq 36$. The approach considers all $\binom{10}{5}^2 = 63,504$ combinations of 5 future X observations, ordered among the 5 data observations from the X sample, and 5 future Y observations, ordered among the 5 data observations from the Y sample. The results for several rank orders are presented in Table 3.14, where the first four rows present cases where the original data lead to rejection of H_0, with all other cases leading to H_0 not being rejected. The NPI lower reproducibility probabilities $\underline{RP}$ are very small when Z is close to 36, substantially smaller than 0.5. This is caused by the same reason as discussed, for a single sample, in Section 3.5, but now the effect is amplified as future observations for both groups are placed at endpoints of their respective intervals such that the overall reproducibility probability is minimized. As we effectively combine lower and upper probabilities for two groups, the resulting combined NPI lower and upper reproducibility probabilities are very imprecise, that is there are big differences between corresponding upper and lower probabilities. When one has larger sample sizes, these differences tend to become smaller, so it reflects the amount of information.

Table 3.14 shows, as in Example 3.2, that $\underline{RP}$ and $\overline{RP}$ are not monotonic in the distance of Z to the critical value 36. In most cases where different ranks per sample lead to the same value of Z, the values of $\underline{RP}$ and $\overline{RP}$ differ, so they depend on the actual ranks per sample and not just on the value of the test statistic (the reported

ranks X-sample	ranks Y-sample	Z	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5	6,7,8,9,10	40	0.25	1
1,2,3,4,6	5,7,8,9,10	39	0.236	0.968
1,2,3,5,8	4,6,7,9,10	36	0.165	0.781
1,2,4,5,7	3,6,8,9,10	36	0.165	0.781
1,2,4,5,8	3,6,7,9,10	35	0.289	0.858
1,2,3,5,9	4,6,7,8,10	35	0.300	0.863
1,2,3,7,8	4,5,6,9,10	34	0.340	0.874
1,2,5,6,9	3,4,7,8,10	32	0.481	0.915
2,3,4,5,9	1,6,7,8,10	32	0.506	0.927
3,4,5,6,9	1,2,7,8,10	28	0.700	0.971
1,2,7,8,10	3,4,5,6,9	27	0.725	0.972
4,5,6,9,10	1,2,3,7,8	21	0.904	0.998
6,7,8,9,10	1,2,3,4,5	15	0.969	1

Table 3.14: NPI-RP for two-sample rank sum test with $H_1 : \delta > 0$, $m = n = 5$, $\alpha = 0.05$, $Z_{0.05} = 36$.

two cases with $Z = 36$ is an exception, which is just due to the same numbers of combinations for which the test result happens to be repeated, it is not a general property). Here it is less clear than in Example 3.2 which further aspects of the data influence the values $\underline{RP}$ and $\overline{RP}$, but it appears that the grouping of neighbouring ranks in a sample influences these values. If $\alpha = 0.004$, H_0 is rejected if $Z \geq 40$ and the results are shown in Table 3.15. The minimum value of $\underline{RP}$ is 0.25 as presented with $\alpha = 0.05$ and the values of $\underline{RP}$ and $\overline{RP}$ depend only on ranks per sample, there are two cases with the same test statistic $Z = 32$, but the bounds of reproducibility probability are different. When X has the largest ranks, the $\underline{RP}$ values increase, and if Y has this property, the $\overline{RP}$ values will increase, but if α is large the increases in these values will be small.

With the same hypotheses and $n = m = 4$, $\alpha = 0.014$ we consider $\binom{8}{4}^2$ orderings, here the H_0 is rejected if $Z \geq 26$. The results of some of these orderings are illustrated in Table 3.16, the importance of ranks in determining the values of $\underline{RP}$ and $\overline{RP}$ appears here also. Although there are some cases that have equal values of test statistic and equal values of $\underline{RP}$ and $\overline{RP}$, but there are other cases that contravene this, for example, if $Z = 24$ the lower and upper bounds of RP are equal, but if $Z = 16$ these values are different. As for the latter case, the same properties of NPI-RP method being with $n = m = 4$ and $\alpha = 0.057$ which is presented in Table 3.17, H_0 is reject if $Z \geq 24$.

ranks X-sample	ranks Y-sample	Z	$\underline{RP}$	$\overline{RP}$
1,2,3,4,5	6,7,8,9,10	40	0.25	1
1,2,3,4,6	5,7,8,9,10	39	0.25	0.827
1,2,3,7,8	4,5,6,9,10	34	0.713	0.948
1,2,5,6,9	3,4,7,8,10	32	0.812	0.970
2,3,4,5,9	1,6,7,8,10	32	0.737	0.950
1,3,5,7,9	2,4,6,8,10	30	0.881	0.984
3,4,5,6,9	1,2,7,8,10	28	0.874	0.981
1,2,7,8,10	3,4,5,6,9	27	0.905	0.987
2,4,6,8,10	1,3,5,7,9	25	0.953	0.995
1,6,7,8,10	2,3,4,5,9	23	0.963	0.997
3,4,7,8,10	1,2,5,6,9	23	0.966	0.997
4,5,6,9,10	1,2,3,7,8	21	0.975	0.998
6,7,8,9,10	1,2,3,4,5	15	0.992	1

Table 3.15: NPI-RP for two-sample rank sum test with $H_1 : \delta > 0$, $m = n = 5$, $\alpha = 0.004$, $Z_{0.004} = 40$.

ranks X-sample	ranks Y-sample	Z	$\underline{RP}$	$\overline{RP}$
1,2,3,4	5,6,7,8	26	0.25	1
1,2,3,5	4,6,7,8	25	0.25	0.832
1,2,3,6	4,5,7,8	24	0.393	0.872
1,2,4,5	3,6,7,8	24	0.393	0.872
1,2,4,6	3,5,7,8	23	0.536	0.913
1,2,5,6	3,4,7,8	22	0.617	0.934
1,2,4,8	3,5,6,7	21	0.636	0.934
1,3,5,7	2,4,6,8	20	0.760	0.966
1,2,5,8	3,4,6,7	20	0.717	0.954
2,4,5,6	1,3,7,8	19	0.758	0.962
1,2,6,8	3,4,5,7	19	0.758	0.962
1,4,5,7	2,3,6,8	19	0.801	0.974
1,2,7,8	3,4,5,6	18	0.774	0.964
1,3,7,8	2,4,5,6	17	0.846	0.981
1,4,6,8	2,3,5,7	17	0.870	0.987
2,4,6,8	1,3,5,7	16	0.899	0.991
1,4,7,8	2,3,5,6	16	0.887	0.989
1,3,6,7	2,4,5,8	16	0.801	0.974
1,5,7,8	2,3,4,6	15	0.907	0.992
1,6,7,8	2,3,4,5	14	0.915	0.993
5,6,7,8	1,2,3,4	10	0.972	1

Table 3.16: NPI-RP for two-sample rank sum test with $H_1 : \delta > 0$, $m = n = 4$, $\alpha = 0.014$, $Z_{0.014} = 26$.

ranks X-sample	ranks Y-sample	Z	$\underline{RP}$	$\overline{RP}$
1,2,3,4	5,6,7,8	26	0.25	1
1,2,3,5	4,6,7,8	25	0.213	0.913
1,2,3,6	4,5,7,8	24	0.172	0.803
1,2,4,5	3,6,7,8	24	0.172	0.803
1,2,4,6	3,5,7,8	23	0.295	0.862
1,2,5,6	3,4,7,8	22	0.389	0.891
1,2,4,8	3,5,6,7	21	0.494	0.914
1,3,5,7	2,4,6,8	20	0.563	0.939
1,2,5,8	3,4,6,7	20	0.568	0.935
2,4,5,6	1,3,7,8	19	0.640	0.953
1,2,6,8	3,4,5,7	19	0.640	0.953
1,4,5,7	2,3,6,8	19	0.632	0.955
1,2,7,8	3,4,5,6	18	0.703	0.964
1,3,7,8	2,4,5,6	17	0.750	0.975
1,4,6,8	2,3,5,7	17	0.750	0.977
2,4,6,8	1,3,5,7	16	0.798	0.985
1,4,7,8	2,3,5,6	16	0.795	0.984
1,3,6,7	2,4,5,8	16	0.632	0.955
1,5,7,8	2,3,4,6	15	0.835	0.990
1,6,7,8	2,3,4,5	14	0.866	0.993
5,6,7,8	1,2,3,4	10	0.952	1

Table 3.17: NPI-RP for two-sample rank sum test with $H_1 : \delta > 0$, $m = n = 4$, $\alpha = 0.057$, $Z_{0.057} = 24$.

We calculated further examples, with these values of m and n and also for other sample size, those results showed similar are presented in Tables 3.18 and 3.19. When we test the same hypotheses with $n = m = 3$ and $\alpha = 0.35$, H_0 is rejected if $Z \geq 12$. There are $\binom{6}{3}^2 = 400$ combinations, some of these combinations are considered in Table 3.18, the first 7 rows show cases which reject H_0 with original data, and the other cases show non rejection of H_0. The $\underline{RP}$ values are small when Z is close to the critical value $Z_{0.35} = 12$, and the minimum value of it is 0.25 whereas the maximum value of $\overline{RP}$ is 1. When Z has equal values, the $\underline{RP}$ and $\overline{RP}$ are similar in some cases and different in others, that is because the $\underline{RP}$ and $\overline{RP}$ do not depend on test statistic only but also on the ranks of the sample. With large ranks of X, the $\underline{RP}$ values are large, and $\overline{RP}$ values are large if Y has large ranks. The same cases of NPI-RP appear with $n = m = 3$ and $\alpha = 0.05$, which are explored in Table 3.19.

However, for larger sample sizes in this method, going through all combinations, becomes quickly computationally infeasible. For example, for $m = n = 7$ the method requires $\binom{14}{7}^2 = 11,778,624$ combinations to be computed to derive the values of $\underline{RP}$ and $\overline{RP}$ corresponding to a single ordering of ranks. Such computations are

ranks X-sample	ranks Y-sample	Z	$\underline{RP}$	$\overline{RP}$
1,2,3	4,5,6	15	0.55	1
1,2,4	3,5,6	14	0.438	0.960
1,2,5	3,4,6	13	0.355	0.900
1,3,4	2,5,6	13	0.355	0.900
1,2,6	3,4,5	12	0.310	0.840
1,3,5	2,4,6	12	0.265	0.828
2,3,4	1,5,6	12	0.310	0.840
1,3,6	2,4,6	11	0.247	0.787
1,4,5	2,3,6	11	0.250	0.797
2,3,5	1,4,6	11	0.247	0.787
1,4,6	2,3,5	10	0.330	0.852
2,3,6	1,4,5	10	0.332	0.840
2,4,5	1,3,6	10	0.330	0.852
1,5,6	2,3,4	9	0.400	0.885
2,4,6	1,3,5	9	0.422	0.910
3,4,5	1,2,6	9	0.5	0.945
3,4,6	1,2,5	8	0.5	0.945
2,5,6	1,3,4	8	0.5	0.945
3,5,6	1,2,4	7	0.580	0.982
4,5,6	1,2,3	6	0.64	1

Table 3.18: NPI-RP for two-sample rank sum test with $H_1 : \delta > 0$, $m = n = 3$, $\alpha = 0.35$, $Z_{0.35} = 12$.

ranks X-sample	ranks Y-sample	Z	$\underline{RP}$	$\overline{RP}$
1,2,3	4,5,6	15	0.25	1
1,2,4	3,5,6	14	0.25	0.840
1,2,5	3,4,6	13	0.400	0.885
1,3,4	2,5,6	13	0.400	0.885
1,2,6	3,4,5	12	0.475	0.900
1,3,5	2,4,6	12	0.550	0.930
2,3,4	1,5,6	12	0.475	0.900
1,3,6	2,4,6	11	0.625	0.945
1,4,5	2,3,6	11	0.640	0.953
2,3,5	1,4,6	11	0.625	0.945
1,4,6	2,3,5	10	0.715	0.968
2,3,6	1,4,5	10	0.700	0.960
2,4,5	1,3,6	10	0.715	0.968
1,5,6	2,3,4	9	0.760	0.975
2,4,6	1,3,5	9	0.790	0.983
3,4,5	1,2,6	9	0.760	0.975
3,4,6	1,2,5	8	0.835	0.990
2,5,6	1,3,4	8	0.835	0.990
3,5,6	1,2,4	7	0.880	0.998
4,5,6	1,2,3	6	0.902	1

Table 3.19: NPI-RP for two-sample rank sum test with $H_1 : \delta > 0$, $m = n = 3$, $\alpha = 0.05$, $Z_{0.05} = 15$.

only required once for a specific case, but for practically interesting data clearly other computational methods are required.

3.7 Concluding Remarks

The NPI approach to reproducibility of tests, as introduced in this chapter, can be extended in several quite obvious ways. For example, one could consider future sample sizes that differ from the data sample size, and one could also explore the use of different levels of significance. Senn [64] discussed the importance of statistical methods that can deal with real-world replications of tests, where circumstances and sample sizes may vary among different tests. However, from the perspective of the theoretical reproducibility of an actual test, and particularly within a frequentist statistical framework, it seems logical to use the same sample sizes and significance levels as in the actual test for which the data are available. It must be emphasized that this use is purely in order to formulate predictive reproducibility in a frequentist context, it does not restrict the applicability in a sense that multiple experiments would have to be stochastically identical copies. While we considered three basic nonparametric tests, the approach is applicable to a wide range of statistical tests, further exploration would provide interesting topics for research where particularly computational aspects may prove to be challenging. For example, it would be interesting to investigate the NPI approach for reproducibility of goodness of fit tests for assumed models, where we expect that a good model fit should lead to quite high values for the NPI lower and upper reproducibility probabilities.

As this chapter has made clear, computational issues are likely to prevent this exact approach, particularly for the rank-based tests in Sections 3.5 and 3.6, to be implemented for practically relevant sample sizes. To resolve this, it is interesting to explore several possibilities. First, there can be substantial benefits in counting more cleverly: we simply went through all combinations, but there is some monotonicity that could be used to reduce this counting effort. Initial investigations showed this to be far from trivial, but nevertheless it provides an interesting topic for research. Secondly, one might be able to find fast methods to derive suitable approximations.

We have not explored this further, but also due to some monotonicity in the counting problems this should be feasible. Thirdly, it is possible to apply a NPI-based bootstrap method such that samples represent the future samples in the exact NPI approach as presented here. This is the route we have taken, and the preliminary results are promising in Chapter 4.

In modern applications of bio-statistics one typically tests many null-hypotheses simultaneously. Reproducibility of such tests is a major challenge, it would be of interest to explore if the NPI approach can make a meaningful contribution to this field. As a further challenge for future research, it would be interesting to go beyond the reproducibility of test results and consider the validation of the statistical models in specific inferences, e.g. prognostic models in medical applications [3]. Such validation is interpreted in terms of the satisfactory performance of the model for future patients, which suggests that a predictive approach like the one presented in this thesis for RP may be suitable.

One aspect that is only mentioned briefly in this thesis is the possibility that one may want to infer on RP given only rejection or non-rejection of the null-hypothesis in the first test. While it is clear that the actual data influence RP substantially, it may be the case that these data are not available. In this situation, the NPI results for set-valued data [18] can be used. For the one-sided one-sample sign test, as presented in Section 3.4, this would lead to the NPI lower probability for the reproducibility of the test conclusion to be equal to the minimum of the NPI lower probabilities for such reproducibility over the possible observation values y which lead to the same conclusion (so rejection or non-rejection), hence it would be equal to 0.5 as proven above. The corresponding NPI upper probability for the reproducibility of the test conclusion is equal to the maximum of the NPI upper probabilities for such reproducibility over the possible observation values y which lead to the same conclusion (so rejection or non-rejection), hence it would be equal to 1. It should be emphasized that these equalities do not generally hold, for example for the corresponding two-sided test the NPI lower and upper probabilities for reproducibility, if only the conclusion of the first test is given, it would need to be derived by a counting argument for the paths in line with the above presentation, but for details

we refer to [18] as we believe it to be only of marginal practical interest.

The NPI approach is strongly based on the data, but as such also dependent on the quality of the data. If, for example, one has reason not to fully trust the data, straightforward application of the methods presented in this thesis would not be appropriate. In such cases, one could perform a sensitivity analysis to check how changes to the data would affect the NPI results, or if one has substantial background information and wishes to use this then a Bayesian approach may be suitable [65]. This raises an interesting further research challenge, namely how NPI methods can be generalized to deal with data that one may not fully trust. It should be emphasized that the use of lower and upper probabilities may help with this challenge, as robustness issues can often be dealt with by going through different possible scenarios (e.g. varying the data in some meaningful way), deriving lower and upper probabilities for each of these, and defining the overall lower and upper probabilities for the inference of interest by taking the minimum of the lower probabilities and the maximum of the upper probabilities corresponding to each of the scenarios, respectively.

Finally, and returning to aspects discussed in the literature and mentioned in Section 1.4, it is important to emphasize that, for the NPI methods presented in this chapter, it is explicitly assumed that future observations are exchangeable with past observations, in the sense of Hill's assumption $A_{(n)}$ on which the NPI approach is based. This assumption is, for example, likely to be violated if the original data resulted from some preliminary selection process, which might occur due to publication bias or other features. This is an important issue which may also lead to interesting further related research challenges, namely exploring whether or not the NPI approach can be developed if the original data are known to have resulted from a preliminary selection process.

Chapter 4

Reproducibility using NPI-Bootstrap

4.1 Introduction

In Chapter 3, we introduced the NPI method for the reproducibility of some non-parametric tests. The limitation of this method is that if the sample size is not very small the computations become complex. To deal with this we use an alternative method for finding the reproducibility probability (RP) of some tests in this chapter, which uses NPI bootstrap method as introduced in Chapter 2. We show the NPI bootstrap approach to finding RP values (NPI-B-RP) with the sign test, one sample signed rank test, two samples rank sum test (these tests were described in Chapter 3 in detail) and the Kolmogorov Smirnov test (KS test) which is shown here in Section 4.5. Note that we did not use this test in Chapter 3 with the NPI-RP method because it is not easy to derive the bounds of NPI-RP for this test. This alternative method can be developed for a variety of tests. We discussed in Section 1.4 the definitions of RP and some different ways to estimate it [29,30,33]. Here we consider the basic idea of RP, which is that the probability for the event that, if the test is repeated based on an experiment performed in the same way as the original experiment, the test outcome (reject H_0 or not) will be the same. We consider RP as prediction not estimation, in Chapter 2 we showed that NPI-B works well for prediction, so for this concept we introduce NPI-B with RP.

We begin this chapter with an explanation of the derivation of reproducibility probability (RP) with NPI-B and nonparametric tests. We apply some nonparametric tests to derive RP using NPI-B in Sections 4.2, 4.3 and 4.4, for one sample sign test, one sample signed rank test (WRS) and two sample rank sum test (WMT), respectively. These tests were explored in Chapter 3 with NPI-RP. In Section 4.5 we present one further test which is the Kolmogorov Smirnov test (KS test) and illustrate the NPI-B method with it. The justification of NPI-B with RP is given in Section 4.6. Several conclusions are given in Section 4.7.

Using the NPI-RP approach to derive the lower and upper bounds of RP needs many calculations and can sometimes be hard with a two sided test, such as the one sample signed rank test and the two samples rank sum test. So we suggest another method for finding the RP using NPI-B. This is based on repeating the experiment a number of times with NPI-B and applying the test to find the proportion of times of the null hypothesis is rejected, or dependent on the outcome of the first experiment. It is the predicted value of RP, NPI-B-RP. We test the accuracy of these estimated values by constructing the confidence intervals for these values. It is an attractive approach because it avoids the complex calculations of the NPI-RP method. We apply NPI-B to find the RP values with a sign test, one sample signed rank test and two samples rank sum test, which are explored in Section 3.2.

4.2 NPI-B-RP for the One Sample Sign Test

One sample sign test (sign test) is one of the basic nonparametric tests and we briefly reviewed it in Section 3.2.1, and used it to find the corresponding NPI-RP in Section 3.4. In this section we want to apply NPI-B-RP to derive the results of the one sample sign test as follows:

1. Draw the original sample X from the specific distribution, then apply the test to find the test statistic K, and the decision of this test.

2. Draw the NPI-B sample from the original one and apply the same test to get the results. Repeat this with 1000 NPI-B samples, then count the number of cases of rejection and non rejection of H_0.

K	frequency	values of NPI-B-RP
3	1	0.845
4	4	0.690 , 0.730 , 0.750 , 0.753
5	6	0.526 , 0.541 , 0.544 , 0.570 , 0.573 ,0.619
6	4	0.593 , 0.605 , 0.616 , 0.667
7	4	0.711 , 0.726 , 0.743 , 0.776
8	4	0.825 , 0.870 , 0.877 , 0.880
9	4	0.930 , 0.957 , 0.957 , 0.967
10	3	0.984 , 0.987 , 0.996

Table 4.1: Sign test, $H_1 : \theta > 2$, $n = 10$, $\alpha = 0.377$, $K_{0.377} = 6$

K	frequency	values of NPI-B-RP
2	4	0.864 , 0.882 , 0.887 , 0.894
3	5	0.698 , 0.730 , 0.739 , 0.750 , 0.778
4	4	0.534 , 0.537 , 0.577 , 0.645
5	4	0.546 , 0.566 , 0.573 , 0.597
6	4	0.713 , 0.723 , 0.756 , 0.760
7	3	0.851 , 0.863 , 0.878
8	5	0.922 , 0.923 , 0.926 , 0.931 , 0.966
9	1	0.973

Table 4.2: Sign test, $H_1 : \theta < 0$, $n = 10$, $\alpha = 0.377$, $K_{0.377} = 4$

3. To find RP, if we reject H_0 in the original experiment, RP is the ratio of the times H_0 is rejected in the 1000 times, or the ratio of the times of non rejection H_0 if we do not reject H_0 in the original experiment.

4. Repeat the previous steps 30 times to record the results.

Table 4.1 shows the results of applying the sign test with the original sample from Gamma (2,2). The hypothesis is $H_0 : \theta = 2$ versus $H_1 : \theta > 2$, and $\alpha = 0.377$ with $n = m = 10$, reject H_0 if $K \geq 6$. The values of α here are typically tabulated values. The parameter θ refers to the median. We can see that the RP values are large with small values of test statistics, like $K = 3$ and $K = 4$, and reduce down with $K = 5, 6$. This means that the minimum value of RP appears when the test statistic approaches the critical value, and then increases again. In Table 4.2, use the sign test with the original sample from Normal (0,1) and test $H_0 : \theta = 0$, $H_1 : \theta < 0$ with $\alpha = 0.377$ and $n = m = 10$. Here H_0 is rejected if $K \leq 4$. The RP values are large with small values and large values of test statistics, and minimize if the test statistic is close to the critical value.

To check if these results follow the same direction of those in Chapter 3, we apply the sign test with the original sample from Normal (0,1) and the same cases of Example 3.1 and Table 3.3 in Section 3.4. However here we repeat the experiment 100 times and record if the NPI-B-RP values are within the corresponding lower and upper NPI-RP results in Example 3.1. For the case $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ and $n = m = 30$ with $\alpha = 0.05$, we found 79% of NPI-B-RP values included in the bounds of NPI-RP. With $\alpha = 0.01$, 85% of these values are within the lower and upper bounds of NPI-RP. For $H_0 : \theta = 0$ versus $H_1 : \theta < 0$, $n = m = 20$, $\alpha = 0.01$ and $\alpha = 0.05$, the ratios are 80% and 86%, respectively. We think that these ratios can be considered as good values because they are representing the most values. This example gives a good picture of NPI-B-RP values, most of these values are located in the corresponding NPI-RP intervals. This leads us to conclude that the results are in line with those in Chapter 3, as we expected. Figures 4.1, 4.2, 4.3 and 4.4 show the values of NPI-B-RP as points. The points which are outside the NPI-RP intervals are presented as $\times$. What is clear from these figures is that these values are few according to other points included in the intervals of NPI-RP.

We considered another illustration of the consistency of the results of the NPI-B-RP of sign test with those in Chapter 3. We used the original sample from Normal (0,1) and $H_0 : \theta = 0$, $H_1 : \theta < 0$ with $n = 10$, $\alpha = 0.377$. H_0 is rejected if $K \leq 4$. We found the NPI-B-RP values and repeated this process 10 times to have 10 values of NPI-B-RP for each value of the test statistic. Then we derived the bounds of NPI-RP for this case, to test if the values of NPI-B-RP are included in the bounds of NPI-RP or not. The results are shown in Table 4.3. We did not show all the values of NPI-B-RP. We simply used the minimum and maximum values. We found that all the values of NPI-B-RP are included in the bounds of NPI-RP. In order to check the performance of the predicted NPI-B-RP values, we consider that these values have a Binomial distribution because they represent rejection or non rejection of H_0. So we use the predicted value NPI-B-RP as the predicted proportion $\hat{p}$ and construct the confidence interval of this value using the formula $\hat{p} \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{1}{n}\hat{p}(1-\hat{p})}$, and then we explore this interval with the bounds of NPI-RP, $[\underline{RP}, \overline{RP}]$. Here we

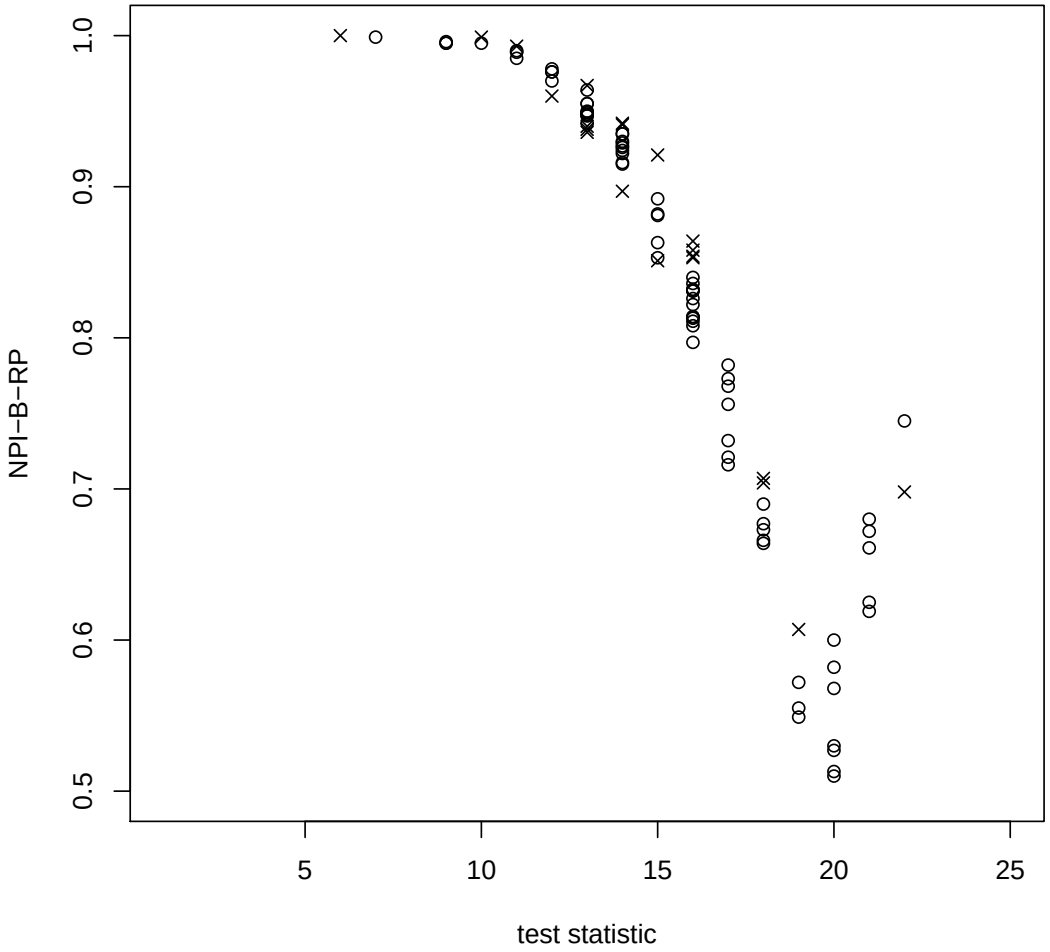

Figure 4.1: Sign test, $H_1 : \theta > 0$, $n = m = 30$, $\alpha = 0.05$

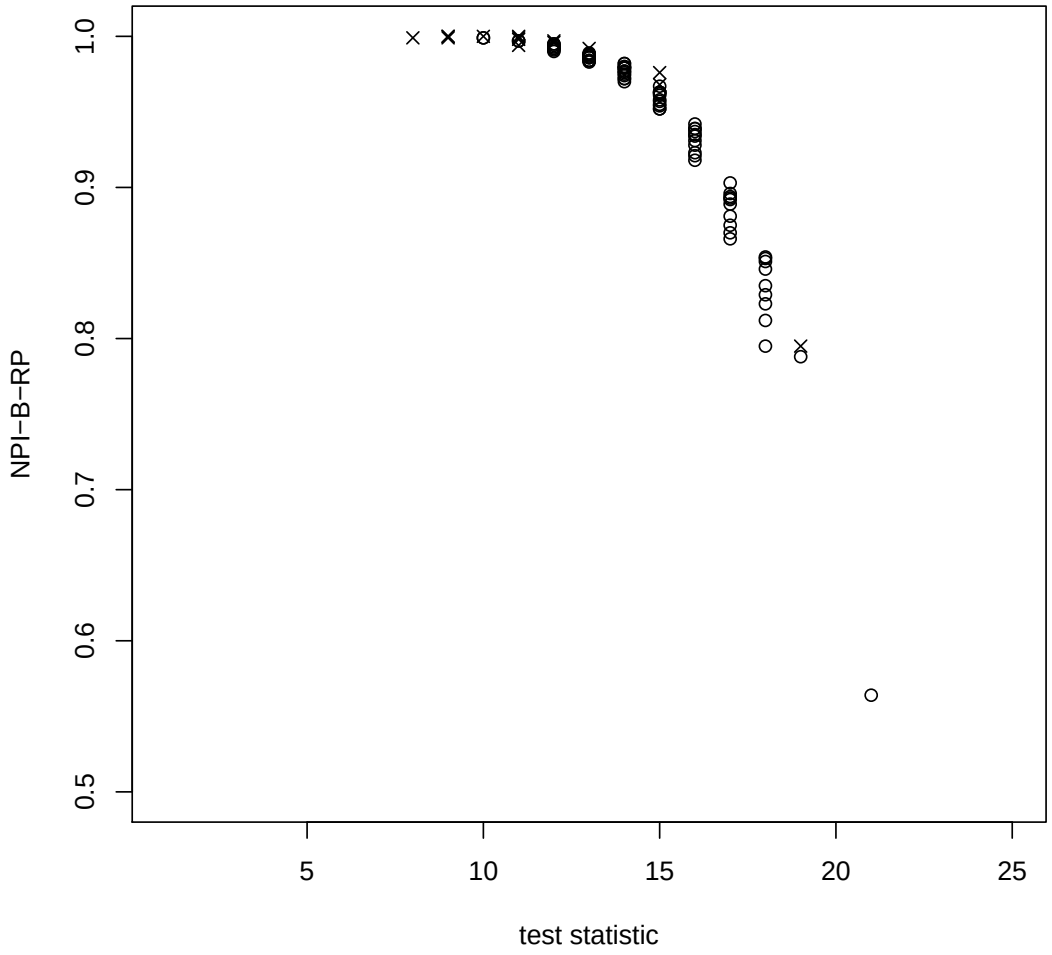

Figure 4.2: Sign test, $H_1 : \theta > 0$, $n = m = 30$, $\alpha = 0.01$

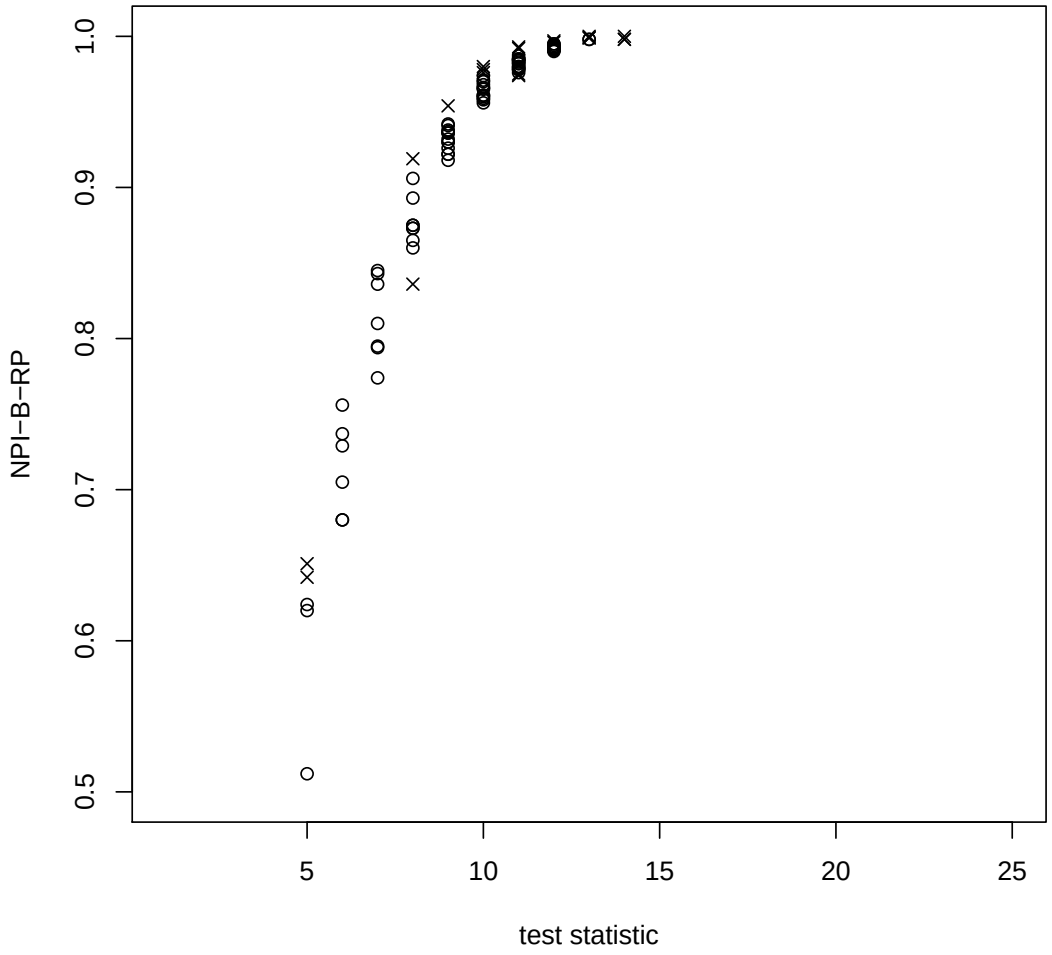

Figure 4.3: Sign test, $H_1 : \theta < 0$, $n = m = 20$, $\alpha = 0.01$

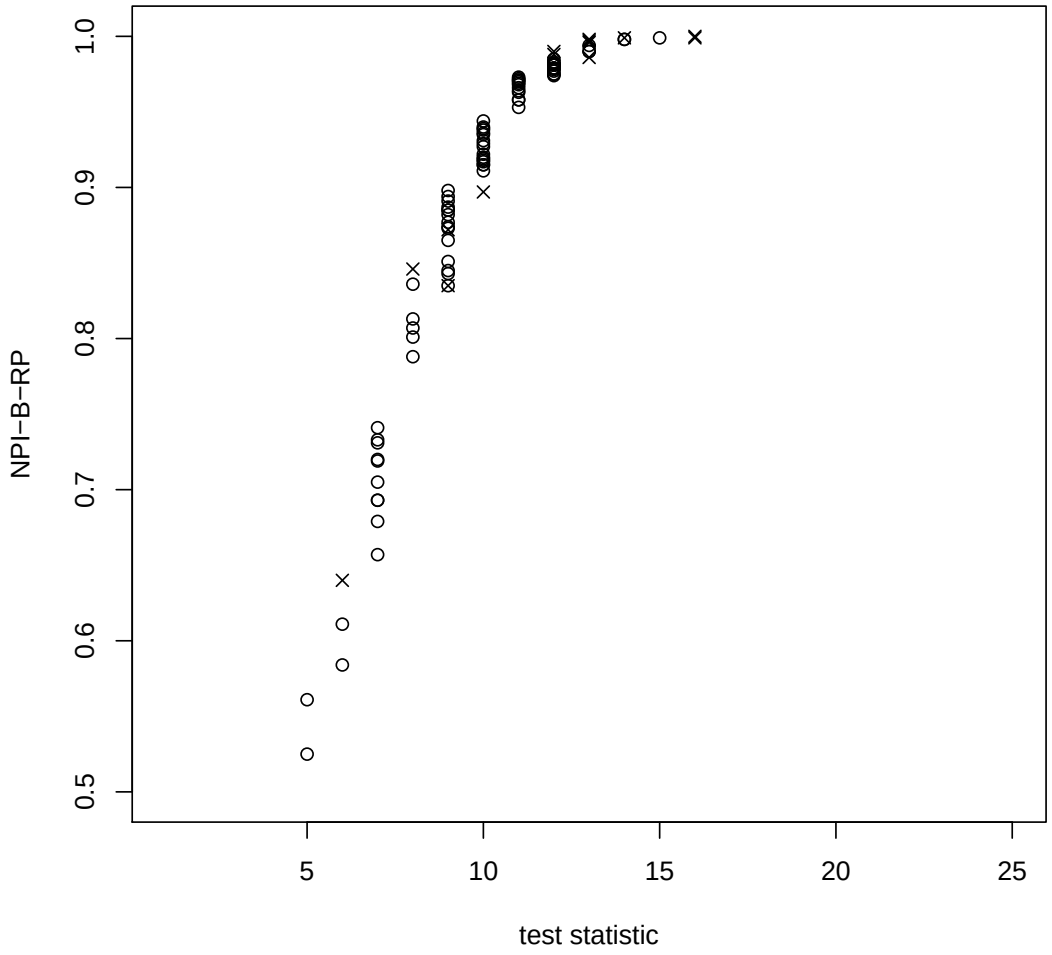

Figure 4.4: Sign test, $H_1 : \theta < 0$, $n = m = 20$, $\alpha = 0.05$

K	$\underline{RP}$	$\overline{RP}$	NPI-B-RP	CI(min)	CI(max)
2	0.825	0.930	0.862,0.899	(0.852,0.872)	(0.891,0.907)
3	0.675	0.825	0.693,0.778	(0.680,0.706)	(0.766,0.790)
4	0.5	0.675	0.518,0.645	(0.504,0.532)	(0.632,0.658)
5	0.5	0.672	0.546,0.682	(0.532,0.560)	(0.669,0.695)**
6	0.672	0.815	0.713,0.813	(0.700,0.726)	(0.802,0.824)*
7	0.815	0.915	0.851,0.914	(0.841,0.861)	(0.906,0.922)*
8	0.915	0.971	0.922,0.975	(0.915,0.929)	(0.971,0.979)*
9	0.971	0.995	0.973,0.996	(0.968,0.978)*	(0.994,0.998)**

Table 4.3: Sign test, $H_1 : \theta < 0$, $n = 10$, $\alpha = 0.377$, $K_{0.377} = 4$

do not compare the confidence intervals of NPI-B-RP with the bounds of NPI-RP, we just want to show that the values of NPI-B-RP will be included in the bounds $[\underline{RP}, \overline{RP}]$ or will be close to them. In Table 4.3 we construct the confidence intervals for the minimum and maximum values of NPI-B-RP (or $\hat{p}$) . This table contains the test statistic K, the bounds of NPI-RP $[\underline{RP}, \overline{RP}]$, the values of NPI-B-RP and the confidence intervals of these two values (CI(min)),(CI(max)). The notation (*) refers to the cases that have NPI-B-RP values included in the bounds of NPI-RP and the CI's overlap with these bounds. The notation (**) refers to cases that have NPI-B-RP values that are not included in the bounds of NPI-RP but the CI's overlap with the bounds $[\underline{RP}, \overline{RP}]$. If the CI's are fully included in $[\underline{RP}, \overline{RP}]$ there is no star. These notations will be used for the other tests in this chapter.

Table 4.3 shows that, when $K = 2, 3, 4$, all of the values of NPI-B-RP are included in $[\underline{RP}, \overline{RP}]$ and the CI's are included in these bounds. When $K = 5$ the minimum value of NPI-B-RP is located in the bounds of NPI-RP and the CI is included in these bounds, but the maximum value does not belong to the bounds of NPI-RP and the CI overlaps with the bounds of NPI-RP. For $K = 9$, the minimum value of NPI-B-RP is included in these bounds and the CI overlaps with them, but the maximum value does not belong within the bounds of NPI-RP and the CI overlaps with them. When $K = 6, 7, 8$ the minimum values of NPI-B-RP are located in the bounds and the CI's are included in them. The maximum values also are located in the bounds of NPI-RP but the CI's overlap with these bounds. These results give a good impression of NPI-B-RP because they show that these predicted values and their confidence intervals are consistent with the bounds of NPI-RP which are shown in Chapter 3. There are no cases where CI's and $[\underline{RP}, \overline{RP}]$ are fully separated.

W	frequency	values of NPI-B-RP	W	frequency	values of NPI-B-RP
14	2	0.809 , 0.815	28	1	0.490
17	1	0.741	29	2	0.536 , 0.539
18	2	0.678 , 0.713	31	1	0.590
20	1	0.644	32	1	0.631
21	1	0.670	33	1	0.594
23	2	0.590 , 0.599	34	1	0.675
24	2	0.552 , 0.571	35	3	0.638 , 0.674 , 0.733
25	1	0.560	39	2	0.743 , 0.764
26	1	0.521	41	1	0.823
27	2	0.517 , 0.581	48	2	0.944 , 0.947

Table 4.4: WRS test, $H_1 : \theta > 0$, $n = 10$, $\alpha = 0.5$, $W_{0.5} = 28$

4.3 NPI-B-RP for the One Sample Signed Rank Test

In Chapter 3 we introduced the one sample signed rank test with an overview about it in Section 3.2.2, and about using it with NPI-RP in Section 3.5. We follow the same steps as in Section 4.2 to find NPI-B-RP but with WRS and Normal (0,1). In Table 4.4 we test $H_0 : \theta = 0$ and $H_1 : \theta > 0$ with $\alpha = 0.5$ and $n = m = 10$. H_0 is rejected if $W \geq 28$. The value of RP is around 0.5 when the test statistic W=27,28,29, and becomes larger when W goes to 14 and when goes to 48. To test $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$ with Normal (0,1) and $n = m = 10$ we use $\alpha/2 = 0.278$. (The tables of Wilcoxon signed rank statistic give the upper bound of probability, so we considered that $\alpha/2 = 0.278$ and then $\alpha = 0.556$ with critical values 21 and 34) and reject H_0 if $W \geq 34$ or $W \leq 21$ with $\alpha = 0.556$. The results show in Table 4.5, the RP values are small when the test statistic is close to the critical values and become larger otherwise.

It is important to know if these results agree with those in Chapter 3. To check this we use the data set $(1, 2, 3, 4, 5, 6)$ of size $n = 6$ from Table 3.8. The test statistic of this data set is $W = 21$ and the lower and upper bounds of NPI-RP are 0.5 and 1, respectively. We resample 1000 NPI-B samples to derive NPI-B-RP value and then repeat this process 100 times to get 100 values of NPI-B-RP. Then we check if these values are considered in the bounds of the NPI-RP approach. We reject H_0 with data set $(1, 2, 3, 4, 5, 6)$, so the NPI-B-RP value is the proportion of the times

W	frequency	values of NPI-B-RP	W	frequency	values of NPI-B-RP
4	1	0.917	30	1	0.382
14	1	0.746	32	1	0.358
16	1	0.726	33	1	0.336
17	1	0.688	34	1	0.666
18	2	0.678 , 0.705	35	2	0.334 , 0.341
19	1	0.676	36	1	0.688
20	1	0.651	37	1	0.686
21	2	0.648 , 0.692	40	1	0.748
22	1	0.352	41	1	0.772
25	2	0.353	46	1	0.868
26	1	0.350	50	1	0.919
28	2	0.346 , 0.363	51	1	0.917
29	1	0.342			

Table 4.5: WRS test, $H_1 : \theta \neq 0$, $n = 10$, $\alpha = 0.556$, $W_{0.556} = 21, 34$

H_0 is rejected in the NPI-B samples. For this case we find all values of NPI-B-RP are within the bounds of NPI-RP. For data set $(-3, -2, -1, 4, 5, 6)$ of size $n = 6$ from Table 3.8, the test statistic is 15 and the lower bound of NPI-RP is 0.728 and the upper bound is 0.902. With this data set we do not reject H_0, so the NPI-B-RP value is the proportion of the times H_0 is not rejected. There are 100% of these values are included in the bounds of NPI-RP.

When we use the data set $(1, 2, 3, 4, 5, 6, 7)$ with $n = 7$ from Table 3.12 we reject H_0, the test statistic is 28, and we find that 100% of the NPI-B-RP values are between the 0.5 and 1, which are the bounds of NPI-RP. If we use the data $(-7, -6, -5, -4, 1, 2, 3)$ from the same table, we do not reject H_0 so the NPI-B-RP is the ratio of cases of non rejection of H_0. All these values are included in the lower bound of NPI-RP 0.939 and the upper bound 0.990. With sample size $n = 4$, we choose data set $(-4, 1, 2, 3)$ which has test statistic $W = 6$ in Table 3.10, and its lower and upper value of NPI- RP are 0.5 and 0.800. In this case we reject H_0 and the NPI-B-RP value is based on the number of times of rejection of H_0, 100% of these values have been in the bounds of NPI-RP. For data set $(-4, -3, 1, 2)$, we do not reject H_0, and the NPI-B-RP depends on the number of cases which do not reject H_0. There are 100% of the NPI-B-RP values are included in 0.457 and 0.757 which are the bounds of NPI-RP.

To consider the confidence intervals of the minimum and maximum values of NPI-B-RP, as we did with the sign test, we use the data sets in Table 3.9. For each set of data we estimate the values of NPI-B-RP and repeat this process 10 times to

sign-ranked data	W	$\underline{RP}(y)$	$\overline{RP}(y)$	NPI-B-RP	CI(min)	CI(max)
1,2,3,4,5,6	21	0.5	1	0.532,0.574	(0.494,0.570)	(0.536,0.612)
-1,2,3,4,5,6	20	0.5	0.773	0.603,0.639	(0.566,0.640)	(0.602,0.676)
-2,1,3,4,5,6	19	0.5	0.773	0.691,0.743	(0.656,0.726)	(0.710,0.776)*
-3,1,2,4,5,6	18	0.5	0.773	0.688,0.739	(0.653,0.723)	(0.706,0.772)
-2,-1,3,4,5,6	18	0.773	0.909	0.805,0.839	(0.775,0.835)	(0.811,0.867)
-4,1,2,3,5,6	17	0.5	0.773	0.704,0.757	(0.670,0.739)	(0.724,0.790)*
-3,-1,2,4,5,6	17	0.773	0.909	0.807,0.852	(0.777,0.837)	(0.825,0.879)
-5,1,2,3,4,6	16	0.5	0.773	0.702,0.767	(0.667,0.737)	(0.735,0.799)*
-3,-2,1,4,5,6	16	0.773	0.909	0.848,0.884	(0.821,0.875)	(0.860,0.908)
-4,-1,2,3,5,6	16	0.773	0.909	0.806,0.859	(0.776,0.836)	(0.832,0.886)
-6,1,2,3,4,5	15	0.5	0.773	0.723,0.768	(0.689,0.757)	(0.736,0.800)*
-3,-2,-1,4,5,6	15	0.910	0.970	0.917,0.941	(0.896,0.938)*	(0.923,0.959)
-6,-3,-1,2,4,5	11	0.910	0.970	0.925,0.954	(0.905,0.945)	(0.938,0.970)
-4,-3,-2,-1,5,6	11	0.970	0.992	0.970,0.983	(0.957,0.983)*	(0.973,0.993)*
-6,-5,-4,-3,1,2	3	0.970	0.992	0.984,0.994	(0.974,0.994)*	(0.988,0.999)**
-6,-5,-4,-2,-1,3	3	0.992	0.999	0.992,0.999	(0.985,0.999)*	(0.997,1)*
-6,-5,-4,-3,-2,-1	0	0.999	1	0.998,0.999	(0.995,1)**	(0.997,1)*

Table 4.6: WRS test, $H_1 : \theta > 0$, $n = 6$, $\alpha = 0.016$, $W_{0.016} = 21$

check whether these values included in $[\underline{RP}, \overline{RP}]$ or not. We find that all the values of NPI-B-RP are included in these bounds, except in the cases which do not belong to the bounds of NPI-RP but the CI's overlap with them (which have **). These cases are shown with the results in Table 4.6. But to apply the NPI-B method with these data sets we need to use the ranks here as the data, and assume that this data are restricted to finite intervals from $-(n+1)$ to $+(n+1)$. That is to apply the main idea of the NPI-B sample. For example, to draw the NPI-B sample from $(1, 2, 3, 4, 5, 6)$ we add $(-7, 7)$ to be $(-7, 1, 2, 3, 4, 5, 6, 7)$ and then resample from it. Table 4.6 explores the CI's of NPI-B-RP values that are included in $[\underline{RP}, \overline{RP}]$ or overlap with them. This shows an agreement with the NPI-RP results in Chapter 3. The same positive picture appears in the data in Table 3.10. We use this data to predict the NPI-B-RP and find their confidence intervals. The results are shown in Table 4.7. If the data set is $(-1, 2, 3, 4)$ we add $(-5, 5)$ to be $(-5, -1, 2, 3, 4, 5)$ and draw the NPI-B sample to predict NPI-B-RP values.

sign-ranked data	W	$\underline{RP}(y)$	$\overline{RP}(y)$	NPI-B-RP	CI(min)	CI(max)
1,2,3,4	10	0.786	1	0.873,0.903	(0.865,0.881)	(0.896,0.901)
-1,2,3,4	9	0.586	0.929	0.753,0.806	(0.742,0.764)	(0.796,0.816)
-2,1,3,4	8	0.543	0.871	0.695,0.713	(0.684,0.706)	(0.702,0.724)
-3,1,2,4	7	0.514	0.829	0.642,0.672	(0.630,0.654)	(0.660,0.684)
-2,-1,3,4	7	0.371	0.757	0.565,0.612	(0.553,0.577)	(0.600,0.624)
-4,1,2,3	6	0.5	0.800	0.585,0.538	(0.573,0.597)	(0.626,0.650)
-3,-1,2,4	6	0.329	0.700	0.507,0.542	(0.495,0.519)	(0.530,0.554)
-4,-1,2,3	5	0.357	0.714	0.514,0.560	(0.502,0.526)	(0.548,0.572)
-3,-2,1,4	5	0.371	0.714	0.522,0.579	(0.510,0.534)	(0.567,0.591)
-4,-2,1,3	4	0.414	0.743	0.595,0.653	(0.583,0.607)	(0.641,0.665)
-3,-2,-1,4	4	0.5	0.843	0.641,0.691	(0.629,0.653)	(0.680,0.702)
-4,-3,1,2	3	0.457	0.757	0.648,0.683	(0.636,0.660)	(0.672,0.694)
-4,-2,-1,3	3	0.557	0.886	0.704,0.756	(0.693,0.715)	(0.745,0.767)
-4,-3,-1,2	2	0.614	0.914	0.752,0.784	(0.741,0.763)	(0.774,0.794)
-4,-3,-2,1	1	0.671	0.929	0.815,0.853	(0.805,0.825)	(0.844,0.862)
-4,-3,-2,-1	0	0.786	1	0.906,0.925	(0.899,0.913)	(0.919,0.931)

Table 4.7: WRS test, $H_1 : \theta > 0$, $n = 4$, $\alpha = 0.438$, $W_{0.438} = 6$

4.4 NPI-B-RP for the Two Sample Rank Sum Test

To apply the two sample rank sum test (Wilcoxon and Mann Whitney test WMT), which was discussed in Sections 3.2.3 and 3.6, and to estimate the RP values with the NPI bootstrap sample we worked through the following steps:

1. Draw two original samples X and Y from any distribution, then apply the WMT test between these samples to get the test statistic Z, and the decision of this test.

2. Draw the NPI-B sample from each original one (X and Y) and apply the WMT test between them to obtain the results. Then repeat this with 1000 NPI-B samples, and then count how many times H_0 is rejected, and how many times H_0 is not rejected.

3. To find NPI-B-RP, if we reject (not reject) H_0 with the two original samples, we need to know how many times H_0 is rejected (is not rejected) in the 1000 NPI-B samples to be the value of RP.

4. Repeat the previous steps 60 times to record the results.

For Table 4.8 the hypothesis is $H_0 : \delta = 0$, $H_1 : \delta > 0$, $\alpha = 0.370$ and $n_1 = n_2 = m = 10$, so $z_\alpha = 110$ and we reject H_0 if $Z \geq z_\alpha = 110$, but here we have

U	frequency	values of NPI-B-RP	U	frequency	values of NPI-B-RP
25	1	0.917	54	2	0.526 , 0.533
27	1	0.917	55	1	0.473
28	1	0.867	56	1	0.531
29	1	0.884	57	2	0.537 , 0.558
33	1	0.853	58	1	0.557
35	1	0.812	59	3	0.553 , 0.570 , 0.600
37	1	0.798	60	3	0.538 , 0.573 , 0.610
40	1	0.787	61	3	0.585 , 0.608 , 0.627
41	1	0.763	63	2	0.642 , 0.656
42	3	0.684 , 0.712 , 0.719	64	1	0.637
43	1	0.677	65	1	0.643
44	1	0.718	66	1	0.683
45	2	0.652 , 0.687	67	1	0.744
47	1	0.659	69	2	0.741 , 0.756
48	1	0.645	70	1	0.750
49	2	0.581 , 0.582	73	3	0.787 , 0.791 , 0.837
50	3	0.529 , 0.550 , 0.638	75	1	0.821
51	2	0.534 , 0.605	77	1	0.814
52	1	0.580	79	1	0.867
53	1	0.555	81	2	0.882 , 0.897

Table 4.8: WMT test, $H_1 : \delta > 0$, $n_1 = n_2 = 10$, $\alpha = 0.370$, $Z_{0.370} = 110$

the Mann Whitney U statistic, as explored in Section 3.2.3, and

$$Z = U + \frac{n_2(n_2 + 1)}{2} \tag{4.1}$$

Then we reject H_0 if $U \geq 55$. We sample the first original sample from Gamma(3,1) and the second one from Gamma(5,2). Here, the critical value is 55 and the RP is 0.473. It is the minimum value of RP in this case and becomes larger with $U = 81, 25$.

In Table 4.9, we test $H_0 : \delta = 0$ and $H_1 : \delta < 0$ with $\alpha = 0.289$ and $n_1 = n_2 = m = 10$. The two original samples are drawn from Uniform(0,1). We reject H_0 if $Z \leq n_1(n_2 + n_1 + 1) - z_\alpha$ or if $U \leq 42$. The value of $U = 42$ appears with the RP value 0.438. It is the minimum value of RP. The maximum value appears with $U = 25$ and with $U = 83$.

In Table 4.10 we have $H_0 : \delta = 0$, $H_1 : \delta > 0$ and $\alpha = 0.370$, $z_\alpha = 110$. The two original samples are drawn from Gamma (5,2). We reject H_0 if $Z \geq z_\alpha = 110$ or if $U \geq 55$. When $U = 55$ the RP is close to 0.5 and becomes larger with extreme values of test statistics.

In Table 4.11, the test is $H_0 : \delta = 0$, $H_1 : \delta < 0$, $\alpha = 0.289$ and $z_\alpha = 113$. We reject H_0 if $Z \leq 97$ or $U \leq 42$. The sample size of the two sample is 10 and the first original sample is drawn from Uniform(0,1) while the second one is from

U	frequency	values of NPI-B-RP	U	frequency	values of NPI-B-RP
18	1	0.896	49	2	0.617 , 0.659
25	1	0.796	50	2	0.614 , 0.648
29	1	0.680	52	2	0.702 , 0.723
31	1	0.718	55	1	0.751
32	3	0.617 , 0.646 , 0.668	56	1	0.760
34	2	0.626 , 0.635	57	3	0.773 , 0.786 , 0.799
35	1	0.601	59	1	0.768
36	1	0.577	60	1	0.796
37	1	0.563	61	1	0.875
38	4	0.519 , 0.538 , 0.541 , 0.541	63	1	0.856
39	2	0.521 , 0.527	64	2	0.850 , 0.882
40	2	0.515 , 0.552	65	1	0.868
41	3	0.494 , 0.496 , 0.528	66	1	0.893
42	1	0.438	68	2	0.906 , 0.916
43	2	0.506 , 0.565	73	2	0.944 , 0.949
44	2	0.563 , 0.583	78	1	0.965
45	1	0.591	82	1	0.976
46	3	0.604 , 0.605 , 0.632	83	1	0.982
47	2	0.618 , 0.658			

Table 4.9: WMT test, $H_1 : \delta < 0$, $n_1 = n_2 = 10$, $\alpha = 0.289$, $Z_{0.289} = 97$

Uniform(0.25,0.5). When $U = 42$ the RP has different values 0.477, 0.491, 0.527, 0.536.

To check if the NPI-B-RP values are in the intervals of NPI-RP, as we did with the last two tests, we use data sets $X = (1, 2, 3, 4, 5), Y = (6, 7, 8, 9, 10)$ and $X = (4, 5, 6, 9, 10), Y = (1, 2, 3, 7, 8)$ from Table 3.14, and $X = (1, 2, 3, 6), Y = (4, 5, 7, 8)$ and $X = (1, 4, 7, 8), Y = (2, 3, 5, 6)$ from Table 3.17. In all these cases, 100% of the NPI-B-RP values are located in the intervals of NPI-RP and that agrees with the results of other tests. To construct the confidence intervals of the predicted values of NPI-B-RP we use the data in Table 3.14. We use the ranks as the data and restrict them in a finite intervals by adding 0 and $(n_1 + n_2 + 1)$ for both X and Y. Then we draw the NPI-B samples to estimate NPI-B-RP and repeat this process many times to test that all these values are included in $[\underline{RP}, \overline{RP}]$. We used the minimum and maximum values to construct the confidence intervals as shown in Table 4.12. The results in this table show once again the agreement between the NPI-B-RP values and the results of Chapter 3. All the values of NPI-B-RP belong to the bounds of NPI-RP and the CI's of these values also belong to these bounds.

U	frequency	values of NPI-B-RP	U	frequency	values of NPI-B-RP
20	1	0.938	54	2	0.504 , 0.507
21	1	0.937	55	2	0.497 , 0.507
25	1	0.913	56	5	0.518 , 0.518 , 0.526 , 0.526 , 0.550
30	3	0.866 , 0.867 , 0.877	57	1	0.523
31	1	0.872	58	3	0.520 , 0.556 , 0.563
33	1	0.859	60	1	0.635
35	1	0.829	61	1	0.609
36	1	0.823	62	2	0.663 , 0.673
37	3	0.749 , 0.763 , 0.770	64	2	0.647 , 0.719
38	2	0.770 , 0.771	65	1	0.689
39	1	0.742	66	2	0.662 , 0.672
41	1	0.769	67	1	0.734
42	1	0.697	68	1	0.704
43	1	0.747	69	1	0.715
45	3	0.663 , 0.674 , 0.693	71	1	0.728
48	1	0.610	73	1	0.770
49	2	0.616 , 0.621	77	1	0.825
50	4	0.581 , 0.596 , 0.606 , 0.628	79	2	0.879 , 0.888
52	1	0.564			

Table 4.10: WMT test, $H_1 : \delta > 0$, $n_1 = n_2 = 10$, $\alpha = 0.370$, $Z_{0.370} = 110$

U	frequency	values of NPI-B-RP	U	frequency	values of NPI-B-RP
27	3	0.711 , 0.751 , 0.811	54	1	0.705
32	2	0.655 , 0.697	56	1	0.731
33	2	0.652 , 0.676	58	2	0.749 , 0.797
36	1	0.607	60	1	0.797
37	1	0.576	61	1	0.779
38	2	0.539 , 0.590	66	1	0.880
40	6	0.523 , 0.530 , 0.534 , 0.555 , 0.560 , 0.590	68	1	0.903
41	5	0.445 , 0.466 , 0.509 , 0.511 , 0.532	69	3	0.879 , 0.900 , 0.917
42	4	0.477 , 0.491 , 0.527 , 0.536	70	1	0.892
43	2	0.512 , 0.540	72	1	0.905
44	1	0.566	73	1	0.929
45	2	0.536 , 0.543	74	1	0.931
47	1	0.589	78	2	0.960 , 0.976
49	1	0.630	80	2	0.961 , 0.965
50	1	0.593	81	1	0.979
51	1	0.718	86	1	0.981
52	1	0.706	89	1	0.995
53	1	0.681	96	1	1

Table 4.11: WMT test, $H_1 : \delta < 0$, $n_1 = n_2 = 10$, $\alpha = 0.289$, $Z_{0.289} = 79$

ranks X-sample	ranks Y-sample	Z	$\underline{RP}(y)$	$\overline{RP}(y)$	NPI-B-RP	CI(min)	CI(max)
1,2,3,4,5	6,7,8,9,10	40	0.25	1	0.640,0.679	(0.610,0.670)	(0.650,0.708)
1,2,3,4,6	5,7,8,9,10	39	0.236	0.968	0.558,0.621	(0.527,0.589)	(0.591,0.651)
1,2,3,5,8	4,6,7,9,10	36	0.165	0.781	0.378,0.431	(0.348,0.408)	(0.400,0.462)
1,2,4,5,7	3,6,8,9,10	36	0.165	0.781	0.384,0.433	(0.354,0.414)	(0.402,0.464)
1,2,4,5,8	3,6,7,9,10	35	0.289	0.858	0.604,0.663	(0.574,0.634)	(0.634,0.692)
1,2,3,5,9	4,6,7,8,10	35	0.300	0.863	0.623,0.663	(0.593,0.653)	(0.634,0.692)
1,2,3,7,8	4,5,6,9,10	34	0.340	0.874	0.649,0.708	(0.619,0.679)	(0.680,0.736)
1,2,5,6,9	3,4,7,8,10	32	0.481	0.915	0.740,0.789	(0.713,0.767)	(0.764,0.814)
2,3,4,5,9	1,6,7,8,10	32	0.506	0.927	0.750,0.803	(0.723,0.777)	(0.778,0.828)
3,4,5,6,9	1,2,7,8,10	28	0.700	0.971	0.875,0.907	(0.855,0.895)	(0.889,0.925)
1,2,7,8,10	3,4,5,6,9	27	0.725	0.972	0.893,0.913	(0.874,0.912)	(0.896,0.930)
4,5,6,9,10	1,2,3,7,8	21	0.904	0.998	0.969,0.984	(0.958,0.980)	(0.976,0.992)
6,7,8,9,10	1,2,3,4,5	15	0.969	1	0.995,0.999	(0.991,0.999)	(0.997,1)

Table 4.12: WMT test, $H_1 : \delta > 0$, $n_1 = n_2 = 5$, $\alpha = 0.05$, $Z_{0.05} = 36$

4.5 NPI-B-RP for the Two Sample Kolmogorov Smirnov Test

The two samples Kolmogorov Smirnov test (KS test) [49,52] is a nonparametric test used to find out if two samples have the same distribution function. Suppose that a first sample is $X_1, X_2, ..., X_{n_x}$ of size n_x has a probability distribution F_x and a second sample is $Y_1, Y_2, ..., Y_{n_y}$ of size n_y has a probability distribution F_y. The two samples are independent and from a continuous population, and the distributions in each one of them are iid. We want to test if there any difference between F_x and F_y, $H_0 : F_x(t) = F_y(t)$ for every t versus $H_1 : F_x(t) \neq F_y(t)$ for at least one t.

Let $\hat{F}_x(t)$ and $\hat{F}_y(t)$ be the empirical distribution functions for X and Y and d be the greatest common divisor of n_x and n_y and set

$$J = \frac{n_x n_y}{d} \max |\hat{F}_x(t) - \hat{F}_y(t)| \tag{4.2}$$

J is the two sided two samples Kolmogorov Smirnov statistic. To compute it let $H_{(1)}, H_{(2)}, ..., H_{(N)}$ the $N = (n_x + n_y)$ ordered values of the combined sample of $X_1, X_2, ..., X_{n_x}$ and $Y_1, Y_2, .., Y_{n_y}$. Now J will be

$$J = \frac{n_x n_y}{d} \max |\hat{F}_x(H_{(i)}) - \hat{F}_y(H_{(i)})| \tag{4.3}$$

At level of significance α reject H_0 if $J \geq j_\alpha$, j_α will be given in tables. If $\min(n_x, n_y) \to \infty$ J is approximately normal [49].

In Table 4.13, we follow the same steps of the WMT test in Section 4.4 to find NPI-B-RP, but with the KS test, to test $H_0 : F_x(t) = F_y(t)$ versus $H_1 : F_x(t) \neq F_y(t)$

J	frequency	values of NPI-B-RP
2	5	0.732 , 0.734 , 0.763 , 0.773 , 0.781
3	4	0.687 , 0.689 , 0.720 , 0.741
4	6	0.600 , 0.620 , 0.624 , 0.630 , 0.671 , 0.693
5	6	0.463 , 0.463 , 0.472 , 0.515 , 0.519 , 0.553
6	5	0.518 , 0.530 , 0.541 , 0.589 , 0.657
7	4	0.674 , 0.735 , 0.770 , 0.774

Table 4.13: KS test, $H_1 : F(t) \neq G(t)$, $n_1 = n_2 = 10$, $\alpha = 0.1678$, $J_{0.1678} = 5$

J	frequency	values of NPI-B-RP
3	3	0.611 , 0.666 , 0.712
4	5	0.409 , 0.409 , 0.451 , 0.507 , 0.555
5	5	0.494 , 0.528 , 0.607 , 0.624 , 0.648
6	5	0.637 , 0.670 , 0.718 , 0.761 , 0.763
7	4	0.766 , 0.794 , 0.803 , 0.815
8	6	0.897 , 0.899 , 0.903 , 0.906 , 0.924 , 0.945
9	1	0.965
10	1	0.980

Table 4.14: KS test, $n_1 = n_2 = 10$, $\alpha = 0.1678$, $J_{0.1678} = 5$

and $\alpha = 0.1678$, $n_1 = n_2 = m = 10$ and resample the two original samples from Normal (0,1). H_0 is rejected if $J \geq 5$. When the test statistic J is close to the critical value, the RP values become small. When $J = 2$ the RP values are large and become small with $J = 5$ and again become larger with $J = 6, 7$. The same situation appears in Table 4.14 when the same hypothesis is tested but draws the two original samples from Uniform (0,1) and Uniform (0.25,0.5), respectively. In the first test we use the same distributions as for the original samples in order to study the case which has the correct null hypothesis, and in the second one we use a different distribution to test the case which has an incorrect null hypothesis. In Table 4.15 we apply the KS test with $\alpha = 0.1678$ and $n_1 = n_2 = m = 10$ with the first original sample from Normal(0,1) and the second from Normal(0,2). In this Table the values of NPI-B-RP become smaller when the test statistic is close to the critical test statistic $J = 5$, and become large with other test statistics.

For all tests when the test statistics are repeated the values of NPI-B-RP are very close. That means there is not a big difference between them. For example, in Table 4.14 when $J = 5$ the values of RP are 0.494, 0.528, 0.607, 0.624, 0.648. These small variations in the results are due to variations in the original samples and in the NPI-B samples.

J	frequency	values of NPI-B-RP
2	5	0.717,0.728,0.746,0.752,0.752
3	8	0.613,0.650,0.664,0.699,0.701,0.702,0.729,0.733
4	9	0.527,0.575,0.596,0.604,0.619,0.626,0.629,0.645,0.664
5	7	0.421,0.422,0.472,0.542,0.542,0.591,0.611
6	1	0.566

Table 4.15: KS test, $n_1 = n_2 = 10$, $\alpha = 0.1678$, $J_{0.1678} = 5$

To develop more general insight into NPI-B-RP values, we implement the four tests: the sign test, WRS test, WMT test and KS test, many times in order to summarize the main properties of NPI-B-RP and plot the boxplot of these values in Figure 4.5. With the KS test the minimum value of the NPI-B-RP values is 0.3780, the maximum is 0.978 and the median is 0.684. For the WMT test these measures are 0.496 , 0.992 , 0.726, respectively. The sign test has a minimum value of NPI-B-RP values it is 0.501 and the maximum is 0.994, while the median is 0.782. With the WRS test the minimum is 0.483, the maximum is 0.967 and the median is 0.674. Note that the distribution of NPI-B-RP values is approximately symmetric in four tests.

4.6 Performance of NPI-B-RP

When a new method is suggested we need to explore some ways of judging the performance of this method. In Sections 4.2, 4.3 and 4.4 we have presented an approach to check if the NPI-B-RP values are included in the NPI-RP intervals or not, we now show two different ways for this. The first method uses the mean square error (MSE) and the other illustrates the predictive inference of the NPI-B method and its power when it is used to find RP values.

4.6.1 Mean Square Error with NPI-B-RP

Using the mean square error (MSE) is a possible way to compare between NPI-B and standard-B as the methods to locate RP values. It is one of the measures of accuracy. To calculate the MSE of RP values we need the estimate of RP, which we already have using NPI-B and standard-B (the details of the results of standard-B

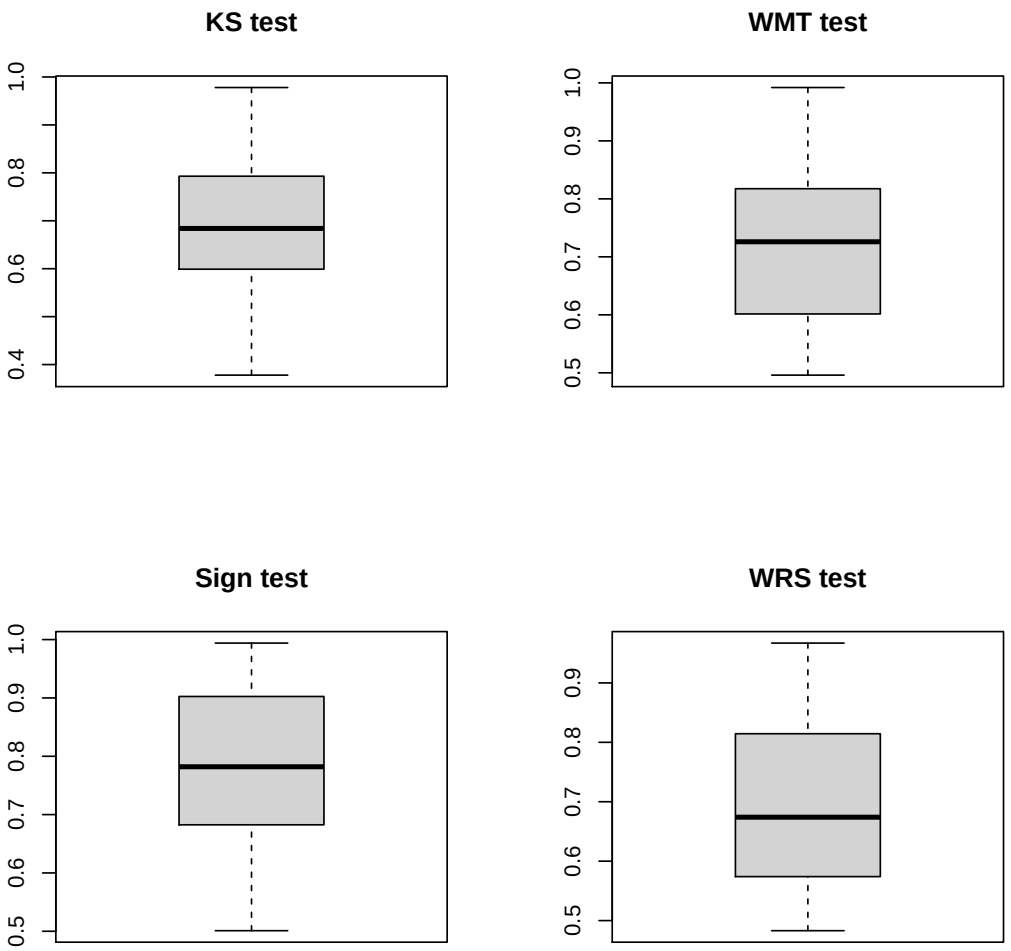

Figure 4.5: Boxplots of NPI-B-RP values from four tests

not consider here). We also need the values of RP. We try to find it by drawing B original samples and applying the test to find RP. For example, the two samples rank sum test draws B pairs of two original samples from Gamma (5,2) with equal sample size 10, and applies the test between these pairs to get the ratio of significant outcomes to be RP1, and the ratio of nonsignificant outcomes to be RP2. We consider RP1 and RP2 as the values of population, because if we repeat this process many times we find approximate values of RP1, such as 0.378, 0.356, 0.375, 0.372 and 0.370. We can deal with one of them as a value of the RP1 and the same status appears with RP2. Note that these two values complement each other.

Now, we follow the same steps which were considered in Sections 4.2, 4.3, 4.4 and 4.5 to arrived at NPI-B-RP1 and NPI-B-RP2 as the estimates of RP1 and RP2 from NPI-B samples, we repeat this V times to have V items of NPI-B-RP1 and NPI-B-RP2, and then find the MSE of these values using this formula

$$MSE = \frac{1}{V}\sum_{v=1}^{V}(estimate_v - parameter_v)^2 \tag{4.4}$$

For the MSE of NPI-B-RP of NPI-B samples the formula will be

$$MSE.NPI.RP = \frac{1}{V}\sum_{v=1}^{V}((NPI.RP1_v - RP1)^2 + (NPI.RP2_v - RP2)^2) \tag{4.5}$$

Then find the MSE for the NPI-B and standard-B (after following the last process to get RP values from standard-B) in order to make a comparison between them in the 10 scenarios which discussed before.

Scenario 1: KS test, Uniform (0,1) and Uniform (0.25,0.5), $\alpha = 0.1678$, $H_1 : F(t) \neq G(t)$.

Scenario 2: KS test, Normal (0,1), $\alpha = 0.1678$, $H_1 : F(t) \neq G(t)$.

Scenario 3: WMT test, Uniform (0,1), $\alpha = 0.289$, $H_1 : \delta < 0$.

Scenario 4: WMT test, Uniform (0,1) and Uniform (0.25,0.5), $\alpha = 0.289$,

$H_1 : \delta < 0$.

Scenario 5: WMT test, Gamma (5,2), $\alpha = 0.370$, $H_1 : \delta > 0$.

Scenario 6: WMT test, Gamma (1,3) and Gamma (5,2), $\alpha = 0.370$, $H_1 : \delta > 0$.

	scenario									
	1	2	3	4	5	6	7	8	9	10
MSE.NPI.RP	0.082	0.181	0.063	0.132	0.052	0.055	0.052	0.040	0.492	0.092
MSE.standard.RP	0.180	0.398	0.106	0.172	0.096	0.094	0.064	0.075	0.541	0.139

Table 4.16: MSE values

	scenario									
	1	2	3	4	5	6	7	8	9	10
mean of NPI-B-RP	0.712	0.641	0.690	0.703	0.691	0.681	0.587	0.663	0.760	0.768
mean of standard-B-RP	0.793	0.718	0.757	0.759	0.767	0.761	0.637	0.733	0.828	0.845
sd of NPI-B-RP	0.172	0.104	0.150	0.173	0.127	0.122	0.209	0.121	0.154	0.143
sd of standard-B-RP	0.285	0.170	0.157	0.175	0.161	0.147	0.212	0.146	0.153	0.144

Table 4.17: summary of RP values

Scenario 7: WRS test, Normal (0,1), $\alpha = 0.556$, $H_1 : \theta \neq 0$.

Scenario 8: WRS test, Normal (0,1), $\alpha = 0.5$, $H_1 : \theta > 0$.

Scenario 9: sign test, Gamma(2,2), $\alpha = 0.377$, $H_1 : \theta > 2$.

Scenario 10: sign test, Normal (0,1), $\alpha = 0.377$, $H_1 : \theta < 0$.

From Table 4.16 we see that using the NPI-B method gives a minimum value of MSE of RP for all scenarios, while the standard-B method has larger values of them. That can be an advantage of the NPI-RP method because it means that the NPI-B method is more accurate when predicting RP. Furthermore, NPI-B gives small values of standard deviation (sd) of RP values in most cases, as Table 4.17 shows. This Table summarises the mean and standard deviation of RP values which are derived from NPI bootstrap and standard bootstrap.

4.6.2 Predictive Performance of NPI-B-RP

It is useful to study a predictive inference with the NPI-B method and there is a possible way to do this as follows:

1. Sample a data set from the selected distribution and apply the test to estimate the NPI-B-RP value.

2. From the same distribution select L further data sets of the same sample size

to consider them as the future samples. Then find the proportion of these samples which have the same test outcome, we refer to this proportion by simulated future proportions (SFP).

3. Repeat this many times (say 100 times) and show the pairs (NPI-B-RP, SFP).

We implemented those steps with some nonparametric tests and plotted the result in Figure 4.6. In this Figure, the upper left plot, we draw the first original sample from Beta (1,2) and the second one from Beta (2,1), then apply the KS test to test $H_0 : F_x(t) = F_y(t)$ versus $H_1 : F_x(t) \neq F_y(t)$ with $\alpha = 0.1678$ and $n_x = n_y = 10$, and plot the pairs (NPI-B-RP , SFP) which are found using the previous steps. We see that 13% of the points (or of the cases) are at the bottom of the plot. Each of these points has a large value of NPI-B-RP and a small value of SFP. That is because the decision of the test with the two original samples was wrong in these cases. But with 87% of the other points at the top, each point has large values of NPI-B-RP and SFP, and these values are close together because the decision taken in the original experiment was correct. From these results we can say that when using NPI-B to predict the RP value, an accurate predictive value is given in 87% of cases. In other words, the SFP values are considered as an indicator of the proportion of the same test outcomes in the future. If the value of NPI-B-RP is close to the value of SFP, that means that the NPI-B is a good method for predicting RP values.

The similar pattern appears in the bottom left plot. In this one we test $H_0 : \theta = 0$ versus $H_1 : \theta > 2$ with a sign test. The original sample is drawn from Gamma (2,2) with size $n = 10$ and $\alpha = 0.377$. Here 8% of the points are in the bottom section of the plot and 92% of them are in the top section. This means that in 92% of the cases, the values of SFP and NPI-B-RP are close. In the right upper plot, the WMT test applied with $H_0 : \delta = 0$ versus $H_1 : \delta < 0$, $\alpha = 0.289$ and $n_1 = n_2 = 10$. The original samples are drawn from Uniform (0,1). In this test 24% of the cases arrived at the wrong decision in the original experiment, and 76% arrived at the correct decision, and with these cases the values of SFP and NPI-B-RP have a similar effect to that in the previous tests. That also happened with the WRS test in 56%, in the right bottom plot, the WRS test the hypothesis $H_0 : \theta = 0$, $H_1 : \theta > 0$ with $\alpha = 0.5$

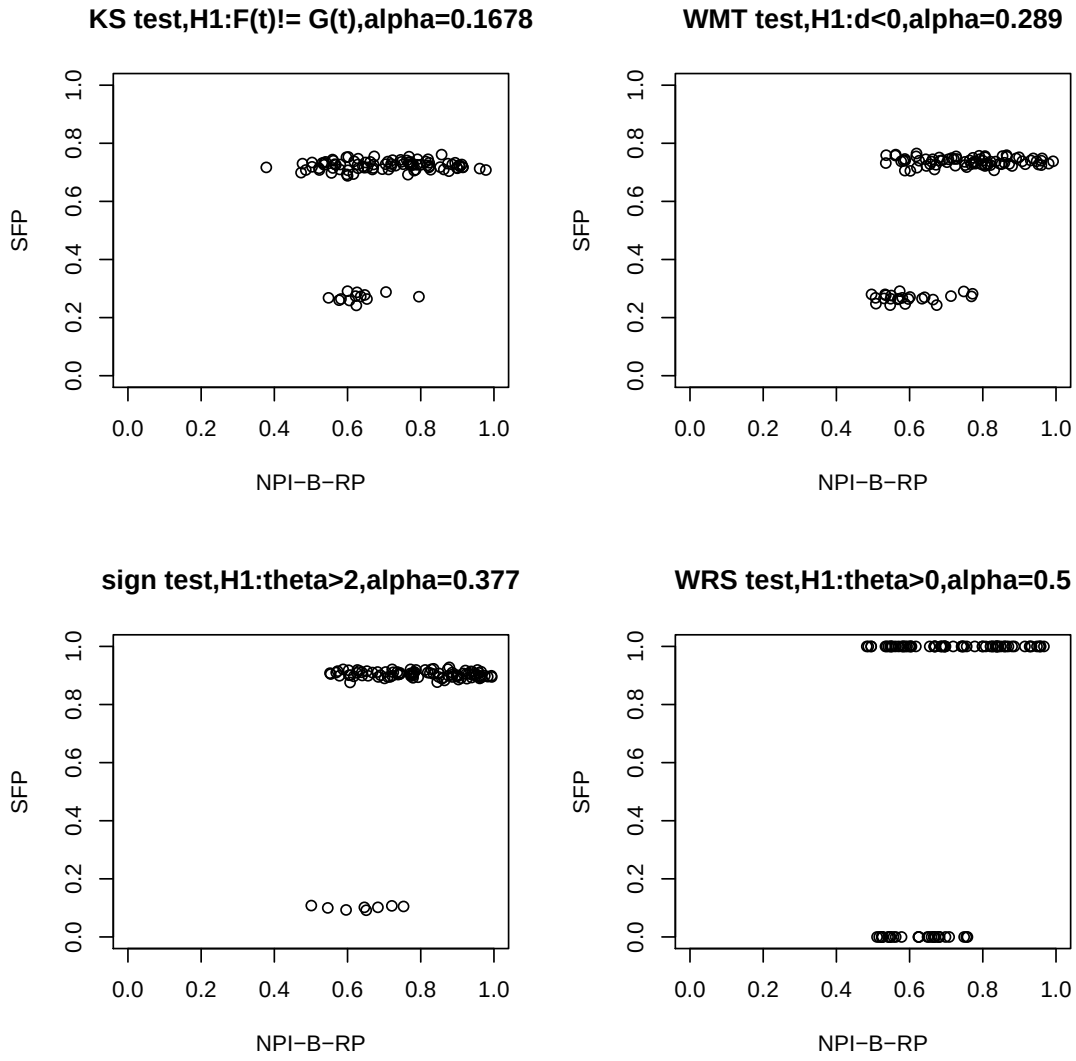

Figure 4.6: Plot the pairs (NPI-B-RP,SFP) of some nonparametric test with sample size 10

and $n = 10$, the original sample from Normal (0,1). The purpose of this section has been to prove the value of the NPI-B approach to predicting RP values and the benefit is clear now because of the strong agreement between SFP and NPI-B-RP values. It is interesting to observe that the results here are in line with those in Chapter 3.

4.7 Concluding Remarks

This chapter explored how the use of NPI-B to estimate RP differs from NPI-RP. It avoids the hardness of calculations and uses the point estimate to present the RP instead of the lower and upper values. Furthermore, it is a flexible approach to use when considering large sample sizes, unlike NPI-RP which is not very flexible. Additionally, the NPI-B-RP values are included in the intervals of NPI-RP and that is consistent with the outcomes in Chapter 3. And to show this consistency in a different way we constructed the confidence intervals for the predicted values of NPI-B-RP and found that all these intervals are included or overlap with the bounds of NPI-RP. Morever, using the NPI-B method makes the variation and MSE of RP values smaller than the standard bootstrap method does, which shows the accuracy and the value of NPI-B method. The simulated future proportions (SFP), which are the proportion of the same test outcomes in the future, are close to the NPI-B-RP values in most cases, that explored the validity of the NPI-B method to predict RP values. This study could be improved using different significance levels and larger sample sizes. This approach can be applied quite straightforwardly and formally to various tests such as the rank correlation test, the Kruskal-Wallis test, tests for two means and tests for two variances [52]. We can resample NPI-B samples from data and apply any such test to find the NPI-B-RP values. This approach will be more complex with tests which use categorical data. It would be challenge to derive the NPI-B-RP for goodness of fit tests. That would be an exciting topic for future research.

Chapter 5

Conclusions

This chapter summarises the main results of this thesis and concludes with a range of ideas for future research. In this thesis we have introduced a new version of bootstrap, nonparametric predictive inference bootstrap (NPI-B), and used it for inference about the reproducibility of tests.

In Chapter 2, NPI-B was presented for data on finite intervals, real line and non negative observations. The NPI-B method goes outside the range of observations, it depends on sampling values from intervals which are created by data values, and adds new values to the data. The NPI-B has more variation than the standard bootstrap. This is a good point of NPI-B because this property eliminates the underestimation of variance which appears in standard-B. The NPI-B does not work well with estimation using confidence intervals, but when its performance with prediction is tested, using prediction intervals, the NPI-B method is a promising alternative to the standard bootstrap for prediction. It often gives good coverage proportions in our study. The results of the NPI-B method with order statistics are consistent with the results of the NPI aspect, and this agreement improves if n is large. It will be interesting, for further research, to study the NPI-B with order statistics using different sample sizes which give a wide picture about its performance from this perspective. We could also use future sample sizes that differ from the data sample.

In Chapter 3, we discussed the NPI method of the reproducibility probability (RP) of some nonparametric tests. RP does not have a unique interpretation in the classical frequentist statistics framework. The NPI lower and upper probabilities are derived for the event that a future test will have the same outcome as the original test. Three nonparametric tests were discussed in this chapter, but we can apply this method with various statistical tests, such as the goodness of fit test. For the sign test we proved that the minimum value of the NPI lower bound of the reproducibility probability is 0.5. For the other tests considered in this chapter no such lower bound for the lower RP has been found and indeed values less than 0.5 occurred. So studying the minimum value of the NPI lower bound of RP with those tests is an idea for future research. Additionally, we considered the two-sided test in the case of the sign test only. Exploring how to derive the bounds of the NPI-RP of the two sided tests of other tests may lead us to further research.

If we consider large sample sizes or more complex hypotheses, the computation is inflexible, so we explored NPI-B as a promising method to fix this in Chapter 4.

In Chapter 4, the NPI-B method was discussed as an alternative method of NPI for predicting the RP of some nonparametric tests. It is a possible method when dealing with large sample sizes or complicated hypotheses. As mentioned before, there is no single definition of RP, but in this chapter we provided the main definition of it. It is the probability that the experiment in the second trial will have the same outcome as the first trial. To contend with this we repeated the experiment many times to sample NPI-B, then found the ratio of times which have the outcome of the first one. It must be referred to that after simulation we found that the predicted values of RP using NPI-B, which are called NPI-B-RP values, are included in the NPI lower and upper probabilities which were explored in Chapter 3. Also the confidence intervals of the NPI-B-RP values are included or sometimes overlap with the bounds of NPI-RP. This means that the results in this chapter are consistent with those in Chapter 3. In addition, the predicted values of RP have less variation and less MSE when NPI-B is used to derive them rather the standard-B. That is a useful attribute of NPI-B. Addotionally the simulated future proportions (SFP),

which are the proportion of the same test outcomes in the future, are close to the NPI-B-RP values in most cases, that explore the validity of the NPI-B method to predict RP values. In Chapter 3 we proved that the minimum value of the sign test using NPI-RP is 0.5. It would be interesting for future research to study what the minimum value of NPI-B-RP values is with different tests.

This study could be extended in different ways, such as using future sample sizes which differ from the data sample size, or different significance levels. Additionally, this method could be applied to different statistical tests such as the goodness of fit test. The NPI-B method works with data on the real line. It is a challenge to know how the NPI-B method works with cells in tables to apply a goodness of fit test using it, to test how well the model fits the specified theoretical distribution and to estimate the reproducibility probability.

To apply the NPI-B method we divided the data set into intervals and sampled the future observation uniformly from this interval, we considered this as a logical assumption. However, on the other hand, it might be not the unique assumption to sample values. Furthermore, to use NPI-B method with data on the real line, we made some assumptions about the distributions of the tails. It would be interesting to think about the effect if we changed these assumptions or used different distributions to sample from the tails or from the intervals. Moreover, in this thesis we used the NPI-B method with data on the real line, is it possible to use it with multivariate data? These suggestions could be the basis for future research into the NPI-B method.

This thesis presented the NPI-B method and its effective performance for prediction. Moreover, the NPI-B method was presented for the reproducibility probability of some nonparametric tests. It is a useful method because it avoids the difficulty of NPI for reproducibility probability and makes the studying of reproducibility more flexible. That is the most important point because the reproducibility of tests is important in the real world, especially in the medical field.

Bibliography

[1] Aboalkhair, A.M. (2012). *Nonparametric predictive inference for system reliability*. PhD Thesis, Durham University, Durham, UK (available from www.npi-statistics.com).

[2] Adekeye, K.S., Lamidi, M.A. and Osanaiye, P.A. (2010). Prediction interval: A tool for monitoring outbreak of some prominent diseases. *Global Journal of Maths and Stat*, **2**, 41–46.

[3] Altman, D.G. and Royston, P. (2000). What do we mean by validating a prognostic model. *Statistics in Medicine*, **19**, 453–473.

[4] Arts, G.R.J. and Coolen, F.P.A. (2008). Two nonparametric predictive control charts. *Journal of Statistical Theory and Practice*, **2**, 499–512.

[5] Arts, G.R.J., Coolen, F.P.A. and van der Laan, P. (2004). Nonparametric predictive inference in statistical process control. *Quality Technology and Quantitative Management*, **1**, 201–216.

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

[7] Augustin, T. and Coolen, F.P.A. (2007). Multinomial nonparametric predictive inference with sub-categories. Proceedings ISIPTA'07, G. De Cooman, L. Vegnaroua, M. Zaffalor , 77–86 .

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

[9] Begley, C.G. and Ellis, L.M. (2012). Raise standards for preclinical cancer research. *Nature*, **483** (29.03.2012), 531–533.

[10] Bickel, P.J. and Freedman, D.A. (1981). Some asymptotic theory for the bootstrap. *The Annals of Statistics*, **9**, 1196–1217.

[11] Binhimd, S. and Coolen, F.P.A (2012). On boostrapping using nonparametric predictive inference. *Proceedings of the 1st ISM International Statistical Conference (ISM 2012)*, 465-468, Malaysia, Department of Mathematical Science, UTM. (www.utm.my/ism-1).

[12] Boos, D.D. and Stefanski, L.A. (2011). P-value precision and reproducibility. *The American Statistician*, **65**, 213–221.

[13] Chernick, M.R. (2008). *Bootstrap Methods: A Guide for Practitioners and Researchers* . Wiley-Interscience.

[14] Collings, B.J. and Hamilton, M.A. (1988). Estimating the power of the two sample Wilcoxon test for location shift. *Biometrics*, **44**, 847–860.

[15] Coolen, F.P.A. (1998). Bayes' postulate repostulated. 6th valencia international meeting on Bayesian statistics, Alcossebre, Spain (May/June 1998) .

[16] Coolen, F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics and Probability Letters*, **36**, 349–357.

[17] Coolen, F.P.A. (2006). On Nonparametric predictive inference and objective Bayesiansm. *Journal of Logic, Language and Information* , **15**, 21–47.

[18] Coolen, F.P.A. (2008). On nonparametric predictive inference for Bernoulli quantities with set-valued data. In: D. Dubois, M. Asuncion Lubiana, H. Prade, M. Angeles Gil, P. Grzegorzewski, O. Hryniewicz (eds.), *Soft Methods for Handling Variability and Imprecision*. Springer, Berlin, 85–91.

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

[20] Coolen, F.P.A and Binhimd, S.. Nonparametric predictive inference for reproducibility of basic nonparametric tests. *Journal of Statistical Theory and Practice*, to appear.

[21] Coolen, F.P.A. and Coolen-Schrijner, P. (2007). Nonparametric predictive comparion of proportions. *Journal of Statistical Planning and Inference*, **137**, 23–33.

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

[23] Coolen , F.P.A and Maturi, T.A. (2010). Nonparametric predictive inference for order statistics of future observations. In: C. Borgelt et al (eds) Combining Soft Computing and Statistical Methods in Data Analusis. Springer, Berlin, 97–104.

[24] Coolen, F.P.A. and Yan, K. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, **129**, 25–54.

[25] Coolen, F.P.A., Troffaes, M.C. and Augustin, T. (2011). Imprecise probability. In: *International Encyclopedia of Statistical Science*, Lovric M. (Ed.). Springer, Berlin, 645–648.

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

[27] Coolen-Schrijner, P. and Coolen, F.P.A. (2004). Adaptive age replacement based on nonparametric predictive inference. *Journal of the Opereational Research Society*, **55**, 1281–1297.

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

[29] Cumming, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, **16**, 1002–1004.

[30] Cumming, G. (2006). Understanding replication: confidence intervals, p values, and what is likely to happen next time. Proceedings of ICOTS-7, 1–6

[31] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and their Applications*. Campridge University Press.

[32] De Capitani, L. and De Martini, D. (2010). Reproducibility probability estimation and testing for the wilcoxon rank sum test. *Rapporto di Ricerca n 191*, Dipartimento di Metodi Quantitativi per le Scienze Economicho ed Aziendali, Universit degli studi di Milano-Bicocca.

[33] De Capitani, L. and De Martini, D. (2011). On stocahstic orderings of the Wilcoxon Rank Sum test statistic with applications to reproducibility probability estimation testing. *Statistics and Probability Letters*, **88**, 937–946.

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

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

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

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

[38] Efron, B. and Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall.

[39] Elkhafifi, F.F. and Coolen, F.P.A. (2012). Nonparametric predictive inference for accuracy of ordinal diagnostic tests. *Journal of Statistical Theory and Practice*, **6**, 681–697.

[40] Elsaeiti, M.A. and Coolen, F.P.A. (2012). A nonparametric predictive approach to sequential acceptance problems. *Journal of Statistical Theory and Practice*, **6**, 383–401.

[41] Gibbons, J.D. and Chakraborti, S. (2011). *Nonparametric Statistical Inference*. Chapman and Hall, Boca Raton.

[42] Good, P.I. (2005). *Resampling Methods: A Practical Guide to Data Analysis*. Birkhauser.

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

[44] Greenwald, A.G., Gonzalez, R., Harris, R.J. and Guthrie, D. (1996). Effect sizes and p values: what should be reported and what should be replicated?. *Psychophysiology*, **33**, 175–183.

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

[46] Hill, B.M. (1988). De Finetti's theoremm induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference. In: Bernardo, J.M., DeGroot, M.H., Lindley, D.V., Smoth, A.F.M. (Eds), *Bayesian Statistics*, **3**, Oxford University Press, Oxford, 211–241 (with discussion).

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

[48] Hjorth, J.S.U. (1994). *Computer Intensive Statistical Methods*. Chapman and Hall.

[49] Hollander, M. and Wolfe, D.A. (1999). *Nonparametric Statistical Methods*. Wiley .

[50] Iverson, G.J., Lee, M.D. and Wagenmakers, E.J. (2010). A model averaging approach to replication: The case of p_{rep}. *Psychological Methods*, **15**, 172–181.

[51] Killen, P.R. (2005). An alternative to null hypothesis significant tests. *Psychological Science*, **16**, 345–353.

[52] Larson, R. and Farber, B. (2006). *Elementry Statistics*, Pearson Prentice Hall.

[53] Lawless, J.F. and Fredette, M. (2005). Frequentist prediction intervals and predictive distributions. *Biometrika*, **92**, 529–542.

[54] 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.

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

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

[57] Maturi, T.A., Coolen-Schrijner, P. and Coolen, F.P.A. (2009). Nonparametric predictive pairwise comparison for real-valued data with terminated tails. *International Journal of Approximate Reasoning*, **51**, 141-150.

[58] Meeden, G. (1993). Noninformative nonparametric Bayesian estimation of quantiles. *Statistics and Probability Letters*, **16**, 103–109.

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

[60] Mojirsheibani, M. (1998). Iterated bootstrap prediction intervals. *Statistica Sinica*, **8**, 489–504.

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

[62] Posavac, E.J. (2002). Using p values to estimate the probability of a statistically significant replication. *Understanding Statistics*, **1**, 101–112.

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

[64] Senn, S. (2002). Comment on 'A comment on replication, p-value and evidence', by S.N.Goodman (Letter to the editor). *Statistics in Medicine*, **21**, 2437–2444. With auther's reply, 2445–2447

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

[66] Sing, K. (1981). On asymptotic accuracy of Efron's bootstrap. *The Annals of Statistics*, **9**, 1187–1195.

[67] Verzani, J. (2005). *Using R for Introductory Statistics*, Chapman and Hall/CRC Press.

[68] Walley, P. (1991). *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, London.

[69] Wilcoxon, F. (1945). Individual comparison by ranking methods. *Biometrics Bulletin*, **1**, 80–83.

[70] Young, G.A. (1994). Bootstrap: More than a stab in the dark? (with discussion). *Statistical Science*, **9**, 382–415.