Parametric Predictive Bootstrap Method for the Reproducibility of Hypothesis Tests

Abdulrahman M. A. Aldawsari[a], Tahani Coolen-Maturi[b,*], Frank P.A. Coolen[b]

[a]*Department of Mathematics, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia*
[b]*Department of Mathematical Sciences, Durham University, Durham, UK*

Abstract

Hypothesis tests are essential tools in applied statistics, but their results can vary when repeated. The reproducibility probability (RP) quantifies the probability of obtaining the same test outcome—either rejecting or not rejecting the null hypothesis—if a hypothesis test is repeated under identical conditions. In this paper, we apply the parametric predictive bootstrap (PP-B) method to evaluate the reproducibility of parametric tests and compare it with the nonparametric predictive bootstrap (NPI-B) method. The explicitly predictive nature of both methods aligns well with the concept of RP. Simulation studies demonstrate that PP-B provides RP values with less variability than NPI-B, benefiting from the assumed parametric model. The bootstrap approach offers a flexible framework for assessing test reproducibility and can be extended to a wide range of parametric tests.

Keywords: Bootstrap, Reproducibility probability, Hypothesis tests, Parametric predictive bootstrap, Nonparametric predictive inference bootstrap

[*] Corresponding author

Email addresses: `abd.aldawsari@psau.edu.sa` (Abdulrahman M. A. Aldawsari), `tahani.maturi@durham.ac.uk` (Tahani Coolen-Maturi), `frank.coolen@durham.ac.uk` (Frank P.A. Coolen)

1. Introduction

The term *reproducible* refers to the ability of results gained from an experiment or statistical analysis of a data set to be reproduced when the study is replicated. Reproducibility is a key concept in scientific methods, providing confidence in knowing exactly what has been achieved. Over recent years, reproducibility has received increasing attention, with several scientific journals launching campaigns to raise awareness of reproducibility issues, such as "Journals Unite for Reproducibility" [35]. Many institutional drug agencies, such as the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMA), require at least two well-controlled clinical trials to evaluate the efficacy and safety of a new drug product before marketing approval [34]. The primary purpose of conducting a second clinical trial is to support the effectiveness of a treatment and to assess whether the clinical results of the first trial can be replicated in the second trial.

Statistical tests serve as tools for experimental evidence to support the effectiveness of a treatment. However, the results of statistical hypothesis tests can vary when the tests are repeated. The concept of reproducibility probability (RP) in the context of hypothesis testing was first addressed by Goodman, who pointed out that there was some misunderstanding about the meaning of the statistical p-value [27]. According to Goodman, the replication probability can illustrate that p-values may exaggerate the evidence against the null hypothesis. In a later discussion, Senn [38] disagreed with Goodman's statement, emphasising the difference between the p-value and RP. However, Senn agreed with Goodman on the importance of reproducibility of test results.

The RP of a test is the probability that the same test outcome, either rejection of the null hypothesis or not, would be reached if the test were repeated based on an experiment performed in the same way as the original experiment. RP indicates the reliability of the result of a statistical hypothesis test. It is particularly relevant when the test leads to the rejection of the null hypothesis, as significant effects in clinical trials may lead to new treatments. For example, Begley and Ellis [5] conducted a study attempting to replicate results from 53 preclinical cancer research studies, confirming the original findings in only 6 cases. Similarly, Bayer HealthCare in Germany found that they could reproduce only about 25% of the results from similar studies. Begley and Ellis emphasised the importance of improving the reproducibility of preclinical studies and building a stronger system, but they did not delve deeply into the statistical techniques used in these studies. They recommended avoiding publication bias toward only positive results and stressed the importance of RP for the reliability of medical tests. These concerns highlight the ongoing challenges in achieving reproducibility, which has led to debates in the literature. The definition and interpretation of RP appear to not be uniquely determined in classical frequentist statistics. For instance, the paper by Simkus et al. [41] addresses variations in the definitions of reproducibility by exploring different experimental contexts, such as changes in datasets, labs, and conditions. It also emphasises the challenges of low reproducibility due to factors like publication bias and poor statistical methods. This work advocates for framing statistical reproducibility as a predictive problem, offering a structured approach to better quantify and address reproducibility challenges.

Recent years have seen growing interest in RP, especially due to its relevance for the practical outcomes of test results. Shao and Chow [39] presented three approaches for evaluating RP in clinical trials: the estimated power approach, the lower confidence bound of power estimates, and the Bayesian approach. These methods estimate the power of a future test using data from previous trials, considering the lower confidence bound as a conservative estimate for RP, especially when the first trial result is highly significant. They argued that a single clinical trial is sufficient if its statistical result is strongly reproducible. De Martini [20] used test power as an estimate for RP in parametric tests and proposed defining statistical tests based on estimated RP. This power-based approach was also followed by De Capitani and De Martini [17, 19, 18] to study RP for nonparametric tests, including the Wilcoxon signed-rank test, sign test, Kendall test, and binomial test. However, the power-based approach is somewhat limited because it only focuses on cases where the null hypothesis is rejected, which is not consistent with the natural interpretation of reproducibility. Additionally, this approach doesn't account for the variability of repeated tests with different data, which is a key factor in understanding RP.

Miller [36] emphasised the importance of distinguishing between two scenarios in test repetition: (1) repetition by independent researchers working under different conditions, and (2) repetition by the same researcher under identical conditions. Miller was sceptical about making precise inferences from an initial test, especially when the true effect size and test power were unknown. In this paper, we focus on the second scenario—repetition by the same researcher under identical conditions—because meaningful frequentist inferences can be derived in this scenario. We define statistical reproducibility for a test as the probability that the same test outcome would be reached if the test were repeated in the same way as the original experiment. We regard assessing test reproducibility as a problem to be solved by predictive inference. It is important to emphasise that we primarily focus on the conclusion of the future test with respect to the null hypothesis based on the actual data of the first test. We do not consider an exact repetition in terms of the same value of the test statistic or the actual observations, nor do we rely solely on the result of the first test as to whether the null hypothesis was rejected or not. Inferring the reproducibility of the test result using actual data seems logical because the strength of the first test's conclusion depends on those data. A prediction of the test result in a future test is more naturally reflected in the final conclusion regarding the rejection or non-rejection of the null hypothesis. We should note that we do not require the sample sizes to be the same for actual and future tests, but this assumption is natural for reproducibility.

This paper employs the recently developed parametric predictive bootstrap (PP-B) method to assess the reproducibility of parametric tests and compares it with the nonparametric predictive bootstrap (NPI-B). Both methods are inherently predictive, considering future observations to form a natural basis for assessing reproducibility (RP), which is framed as a problem of prediction rather than estimation. The terms PP-B-RP and NPI-B-RP refer to the reproducibility values derived from the PP-B and NPI-B methods, respectively, and are discussed in Section 3. Before introducing these methods, Section 2 provides an overview of a predictive approach to statistical reproducibility, first introduced by Coolen and Binhimd [12] within the framework of nonparametric predictive inference (NPI). This section highlights the advantages of viewing reproducibility as a predictive problem rather than an estimation problem and discusses key challenges associated with traditional approaches to reproducibility assessment. The paper then explores the use of PP-B for parametric tests, comparing it with NPI-B, which also employs predictive bootstrap techniques, in Sections 4, 5, and 6. Section 7 compares these two methods to the traditional NPI-RP method in the context of the likelihood ratio test. The paper concludes in Section 8.

2. Statistical Reproducibility: A New Perspective and Challenges

A new perspective on test reproducibility was introduced by Coolen and Binhimd [12] within the framework of nonparametric predictive inference (NPI), a frequentist statistical method. They applied the NPI approach to assess reproducibility probability (NPI-RP) for a variety of nonparametric tests, including the sign test, Wilcoxon's signed-rank test, and the two-sample rank-sum test. This method uses the test result for a predicted future sample that is the same size as the original sample, which reflects the essence of reproducibility. The NPI approach is explicitly predictive, focusing on future observations, and makes minimal assumptions about the data, which leads to imprecision that can be quantified by the use of lower and upper probabilities. The NPI-RP framework views reproducibility from the perspective of prediction rather than estimation, which sets it apart from the more traditional power-based approach to reproducibility. This framework offers reproducibility probabilities for both the rejection and non-rejection of the null hypothesis, which is significant since much of the focus is typically on tests that lead to rejection, especially in fields like clinical trials where significant effects often lead to new treatments. However, we believe that reproducibility should also be considered for tests that do not reject the null hypothesis in order to provide a more complete picture of test reliability. The NPI approach has been extended to other nonparametric tests, including the quantile test and the precedence test [11].

The core idea of the NPI-RP approach is to consider all possible orderings of future observations among the data observations. It takes into account the different ways that future data could be arranged among the original data, with each arrangement having an equal chance of occurring. These future observations are grouped into intervals, and while we don't know the exact values of the future data for each possible ordering, we can predict how many observations will fall within each interval. Importantly, there are no further assumptions placed on the future data—each data point can take any value within its designated interval. By examining all possible arrangements of future data, the NPI-RP approach allows us to compare the conclusions of tests applied to these future datasets with the conclusion of the original test. The proportion of future tests that lead to the same conclusion as the original test is then used to determine the reproducibility probability.

However, the NPI-RP approach becomes computationally expensive for large datasets. For example, with just a sample size of 15, the number of possible future arrangements of the data can become prohibitively large, requiring the calculation of an enormous number of potential orderings to derive reproducibility probabilities. To address this issue, Coolen and Marques [14] proposed a sampling methodology. Instead of calculating every possible ordering, they suggest randomly sampling future data arrangements. This method satisfies the conditions of simple random sampling (SRS), where each future arrangement has an equal probability of being selected, and each selection is independent of the others. By using a sufficiently large number of samples, the differences between sampling with and without replacement become negligible, making this approach computationally feasible. The sampling process involves selecting a vector of integers that corresponds to the ranks of the ordered data observations. The future data is then simulated based on these sampled ranks, which allows the reproducibility probability to be estimated without the need to examine all possible arrangements.

Another approach to improving computational efficiency was proposed by Coolen and Binhimd [13], who introduced an NPI-based bootstrap method. This method estimates reproducibility probability by generating future data samples through resampling, thereby simplifying the calculation of reproducibility values for various nonparametric tests. Further details on this bootstrap method will be discussed in Section 3.3.

3. Bootstrap methods

3.1. Classical bootstrap methods

The bootstrap method, introduced by Bradley Efron in 1977 and detailed in a 1979 Annals of Statistics paper [8, 23], uses resampling techniques to quantify uncertainty in sample estimates. Known for its straightforward implementation and effectiveness, it provides researchers with a valuable alternative to complex derivations when no analytical solution is available [25]. By leveraging computational power, the bootstrap method assesses the statistical accuracy of complex procedures and is widely used for hypothesis testing due to its simplicity. Chernick [7] discusses applications of the method in areas such as hypothesis testing, confidence intervals, regression, and time series analysis. Various adaptations, including double, smooth, and Bayesian versions, have also emerged [4, 16, 40]. These methods apply to diverse data types, such as real [29], right-censored [1], and ordinal data [6].

The standard bootstrap method, introduced by Efron [25], is a nonparametric approach that resamples from the original data set to quantify uncertainty in sample estimates. Efron's Bootstrap (EB) involves repeatedly resampling with replacement from the original observations, giving each observation an equal chance of being selected during the resampling process [32]. With minimal mathematical assumptions, the EB method is easy to implement using statistical software, making it highly popular in applied statistics. Importantly, EB does not assume any specific data distribution [28, 37]. In contrast, parametric bootstrap (PB) assumes that the data follow a known distribution with unknown parameters. This method involves drawing samples from the assumed distribution using estimated parameters rather than resampling from the original data. The main idea of the PB method is to estimate the parameters of the presumed distribution based on the observed data, then generate multiple PB samples from this distribution using the estimated parameters [28, 33]. While this approach can include observations not present in the original sample, it requires knowledge of the data distribution; if the assumed model is incorrect, results may be misleading. Unlike PB, the EB method makes no distributional assumptions and includes all observations from the original sample, with ties in the data being preserved. Thus, PB is more suitable when there is prior knowledge of the population's distribution.

The rest of the section describes two bootstrap methods: parametric predictive bootstrap (PP-B) [2] and nonparametric predictive inference bootstrap (NPI-B) [13]. Both focus on predictive inference, but NPI-B does not assume a specific distribution for the data, whereas PP-B requires distributional assumptions.

3.2. Parametric predictive bootstrap (PP-B)

This section provides a brief overview of the basic concept of the PP-B method for real-valued data. For details about the method and its implementation, we refer the reader to Aldawsari [2]. The PP-B method involves sampling a single observation from an assumed distribution with estimated parameters based on an original data set of size n. This observation is then added to the data, and the process is repeated with $n + 1$ observations. In order to sample the second observation, we re-estimate the distribution parameters with the new observation added to the data. Continuing this process to sample m further values in the same way, each observation adding to the data and re-estimating the parameters before sampling the next one. The PP-B includes only the m sampled observations, so it excludes the n original data observations. The PP-B algorithm for one-dimensional real-valued data is as follows:

1. We have a random sample consisting of n observations $x_1, x_2, \ldots, x_n$ from a known distribution $F(x;\theta)$, with parameter θ.
2. The parameter θ of the assumed distribution is estimated by $\hat{\theta}$ from the available data, using maximum likelihood estimation (MLE) or any other estimation method.
3. Sample one future observation x_1^* randomly from the fitted distribution $F(x;\hat{\theta})$.
4. Add x_1^* to the data giving data set $(x_1, x_2, \ldots, x_n, x_1^*)$; increase n to $n+1$.
5. Repeat Steps 2-4, now with $n+1$ data, to obtain a further future value. This process continues until m observations have been sampled in total, with each one added to the data and the parameters re-estimated before sampling the next observation. These sampled observations $x_1^*, x_2^*, \ldots, x_m^*$ form a PP-B sample of size m.
6. Repeat Steps 2-5 to obtain B of PP-B samples of size m.

3.3. Nonparametric predictive inference bootstrap (NPI-B)

Coolen and Binhimd [13] presented a nonparametric predictive bootstrap technique rooted in a frequentist approach known as NPI. The NPI (nonparametric predictive inference) method has evolved over the last two decades to address various applications and statistical challenges involving different types of data. NPI is a statistical technique based on Hill's assumption $A_{(n)}$ that makes inferences on a future observation based on past data observations [9, 10]. Hill [30] introduced the assumption $A_{(n)}$ for prediction of one future observation X_{n+1} with no prior knowledge about the underlying distribution. Suppose that $x_1, \ldots, x_n$ are the observed data corresponding to real-valued and exchangeable random quantities $X_1, \ldots, X_n$. Let $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$ be the ordered observations and define $x_{(0)} = -\infty$ and $x_{(n+1)} = +\infty$ for ease of notation. The assumption $A_{(n)}$ states that the future observation X_{n+1} is equally likely to fall in any open interval $(x_{(i-1)}, x_{(i)})$, $i = 1, \ldots, n+1$. These intervals were created by the previous n observations between consecutive order statistics of the given sample.

The assumption $A_{(n)}$ itself is not sufficient to derive precise probabilities for any event of interest, but it can be used to derive bounds (lower and upper) of probabilities, which are called imprecise probabilities. The NPI approach is introduced by Coolen and Augustin [3, 15] which uses lower and upper probabilities for events of interest considering future observations based on Hill's assumption. The lower probability is the maximum lower bound for the precise probability for the event and denoted by $\underline{P}(\cdot)$. The upper probability is the minimum upper bound for the event and denoted by $\overline{P}(\cdot)$.

Sequential application of the assumptions $A_{(n)}, \ldots, A_{(n+m-1)}$ can be used to generalise NPI for $m \geq 1$ future real-valued observations based on n real data observations. These assumptions imply that all $\binom{n+m}{n}$ possible different orderings of the m future observations among the n data observations are equally likely to appear, with no further assumptions made on where future observations will be within any of these intervals $(x_{(i-1)}, x_{(i)})$ [14].

Coolen and Binhimd [13] introduced a predictive bootstrap method based on NPI, called nonparametric predictive inference bootstrap (NPI-B). The NPI-B method involves creating $n+1$ intervals between the n ordered observations of the original data and then selecting one of these intervals randomly. The first observation is drawn uniformly from the selected interval, which is then added to the original data, resulting in $n+1$ observations. This creates a partition consisting of $n+2$ intervals, from which the second observation is sampled. The process continues until m observations are drawn, where m is predefined. These m observations constitute one NPI-B sample (which, of course, does not include the n original data observations). In NPI-B, all possible orderings of the new observations among the past observations are equally likely to occur. NPI-B's sampling method, which involves drawing each observation from the intervals in the partition created by combining the n original observations together with all previously drawn observations belonging to the same bootstrap sample, leads to more variation in bootstrap samples than Efron and parametric bootstrap samples.

It is worth mentioning that one observation is sampled uniformly from each chosen interval when applying NPI-B. However, it cannot be sampled uniformly from an open-ended interval, e.g., data defined on the whole real line lead to the first and last intervals in the form of $(-\infty, x_{(1)})$ and $(x_{(n)}, +\infty)$. Coolen and Binhimd [13] suggest using the tail of a Normal distribution for real-valued data and the tail of an Exponential distribution for non-negative real-valued data. It is important to note that the conditional tail distribution is only used to sample an observation from open-ended intervals; otherwise, the observation is sampled uniformly from finite intervals. The NPI-B algorithm for real-valued data on finite and infinite intervals is as follows:

1. Create $n+1$ intervals between the n ordered observations $(x_{(0)}, x_{(1)})$, $(x_{(1)}, x_{(2)}), \ldots, (x_{(n-1)}, x_{(n)}), (x_{(n)}, x_{(n+1)})$, where $x_{(0)}$ and $x_{(n+1)}$ are the end points of the possible data range.
2. Select one of the $n+1$ intervals randomly, each with equal probability, and sample one future observation uniformly from this selected interval.
   (a) We sample the future value uniformly for any finite interval.
   (b) For the case with data on the whole real line $(-\infty, +\infty)$: If the chosen interval is $(-\infty, x_{(1)})$ or $(x_{(n)}, +\infty)$, we sample the future value from the tail of Normal distribution with mean $\mu = \frac{x_{(1)}+x_{(n)}}{2}$ and standard deviation $\sigma = \frac{x_{(n)}-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ^{-1} indicates the inverse function of a standard normal cumulative distribution function.
   (c) For the case with data on the $(0, +\infty)$: If the chosen interval is $(x_{(n)}, +\infty)$, we sample the future value from the tail of Exponential distribution with rate $\lambda = \frac{\ln(n+1)}{x_{(n)}}$.
3. Add this sampled observation x_1^* to the data; increase n to $n+1$.
4. Repeat Steps 1-3, now with $n+1$ data, to obtain a further future value. This is continued to sample m future observations from the intervals in the partition created by combining the n original observations with all previously drawn observations that belong to the bootstrap sample. These m drawn observations $(x_1^*, x_2^*, \ldots, x_m^*)$ form one NPI-B sample of size m.
5. Repeat Steps 2-4 to obtain B of NPI-B samples of size m.

3.4. Classical vs predictive bootstrap methods

The method for sampling observations in NPI-B, where each observation is drawn from the intervals created by combining the n original observations with all previously drawn observations belonging to the same bootstrap sample, results in more variation in bootstrap samples than in EB and PB. PP-B's sampling method, which adds the sampled observations to the data set and estimates the parameter before sampling the next observation, also causes more variation in the bootstrap samples than the EB and PB samples. All observations are sampled based on the original data only in the EB and PB methods. The EB method relies on a resampling process with replacements from the original data set, where each value of the original data set has the same probability of being selected by random during the resampling process [24]. In the PB method, the data are assumed to come from a known distribution with unknown parameters. The parameters of the assumed distribution are estimated using the available data, then observations are sampled from the assumed distribution with the estimated parameters to obtain PB sample [28]. Bootstrap samples in PP-B, NPI-B, and PB do not restrict themselves to already observed values, whereas in EB samples, all observations are included in the original sample.

4. Bootstrap-RP for the one-sample t-test

The one-sample t-test is a statistical test used to determine if the mean of a population differs from a specified value. Given a random sample $X_1, X_2, \ldots, X_n \sim N(\mu, \sigma^2)$ from a normal population with unknown variance σ^2, the hypotheses of interest are $H_0 : \mu = \mu_0$ against $H_a : \mu \neq \mu_0$, $\mu > \mu_0$, or $\mu < \mu_0$, depending on the test direction [26]. If the sample is normally distributed, the test statistic under the null hypothesis is:

$$T = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

where t_{n-1} is the t-distribution with $n-1$ degrees of freedom, and $\bar{x}$ are s^2 are the sample mean and variance. The null hypothesis H_0 is rejected at significance level α in favour of the two-sided alternative $H_a : \mu \neq \mu_0$ if $|T| > t_{n-1}^{(1-\alpha/2)}$, where $t_{n-1}^{(1-\alpha/2)}$ is the $(1-\alpha/2)$ percentile of the t-distribution with $n-1$ degrees of freedom. For a one-sided upper-tailed test $H_a : \mu > \mu_0$, we reject H_0 if $T > t_{n-1}^{(1-\alpha)}$; for a one-sided lower-tailed test $H_a : \mu < \mu_0$, we reject H_0 if $T < t_{n-1}^{(\alpha)}$.

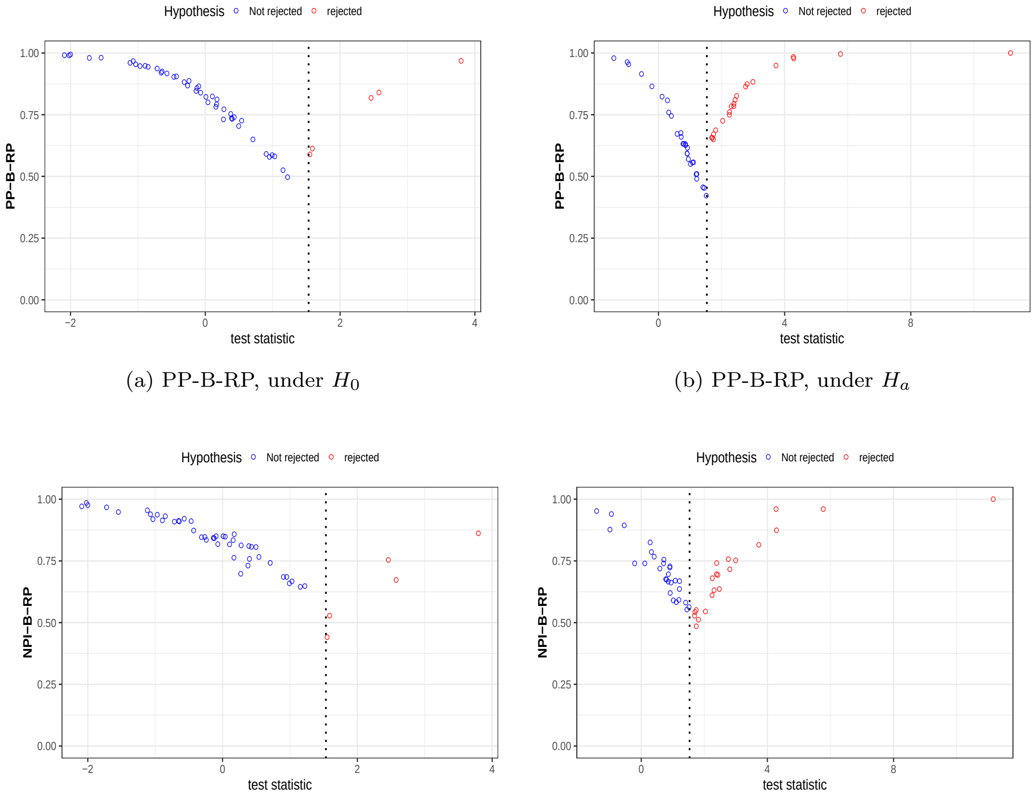
This section studies the reproducibility probability (RP) of the one-sample t-test using the bootstrap method. We apply both the PP-B and NPI-B methods to assess RP and compare their performance. Since test reproducibility is a predictive inference problem, the explicitly predictive nature of these methods provides an appropriate framework for inferring RP. Simulation studies are conducted to compare the two bootstrap methods for evaluating the RP of the one-sample t-test, as follows:

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

2. Draw a bootstrap sample of size n from the sample X and apply the same test to obtain the decision of this test.
3. Perform Step 2 in total B times and record the test result each time whether the null hypothesis is rejected or not.
4. The estimate of the RP is the ratio of B times in which the original sample and the bootstrap samples have the same conclusion.
5. Perform all these steps N times to obtain RP values for both rejection and non-rejection cases of the null hypothesis.

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

We first examine the relationship between Bootstrap-RP and the test statistic for the one-sample t-test in the simulations. The values of RP for the two methods tend to increase when the test statistic moves away from the test thresholds, as expected, regardless of the decision on H_0. The worst-case scenario gives an RP of about 0.5 when the original test statistic is close to the test threshold. Without further information, one would expect a repeat experiment to produce a second test statistic whose value is equally likely to be larger or smaller than the original test statistic, and therefore, the same conclusion would be reached with a probability of 0.5. A repetition of an experiment that had an original test statistic far away from the test threshold is likely to produce a second test statistic that is also far away from the test threshold. Therefore, the RP values tend to increase when the test statistic moves away from the test thresholds. Simulation studies show that RP values based on PP-B have less variability than NPI-B because of the parametric model assumed for PP-B. There is a clear fluctuation observed in the values of RP based on NPI-B because this bootstrap method does



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

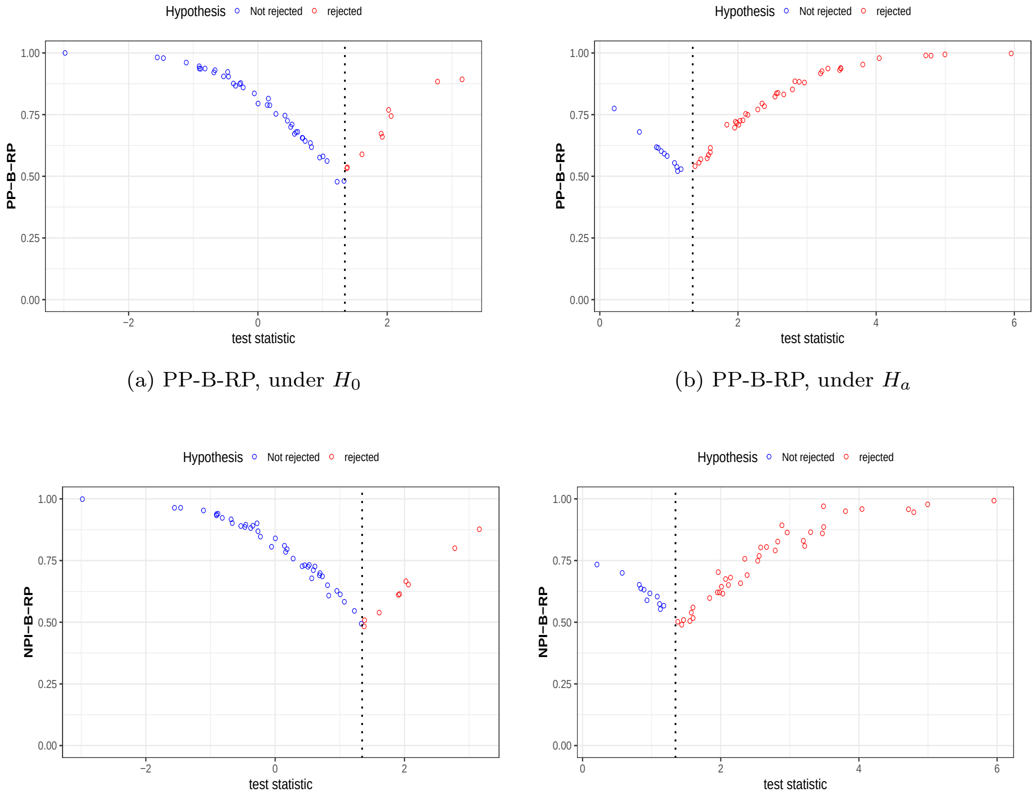
(d) NPI-B-RP, under H_a

Figure 1: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for one-sample t-test, where $n = 5$.

not assume a parametric model, and the sample size is quite small. The fluctuation of RP values based on NPI-B is more visible when simulations are conducted under H_a due to more cases of test statistics close to the test threshold.

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

Additionally, we analyse the impact of increasing sample size on the patterns of Bootstrap-RP values. The results of RP values based on the two bootstrap methods for samples of size $n = 15$ under H_0 and H_a are presented in Figure 2. As the sample size increases, the Bootstrap-RP value becomes closer to 0.5 when the observed test statistics are close to the test threshold in both cases of rejection and non-rejection. Also, the fluctuation in NPI-B-RP values is decreased when the sample size increases. The power of the test is positively correlated with sample size, which means a larger sample size gives greater power. It is because a larger sample size narrows the distribution of the test statistic, so the false null hypothesis can be distinguished more clearly from the true null hypothesis. For simulations under H_a, increasing sample size leads to more cases rejecting H_0, which simply results from the test becoming more powerful with a larger sample size. The



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_a

Figure 2: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for one-sample t-test, where $n = 15$.

(a) Under H_0

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

(b) Under H_a

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

Table 1: *Simulation under H_0 and H_a: values of RP of one-sample t-test using PP-B and NPI-B methods with four observed samples of sizes $n = 5$ and $n = 15$.*

pattern of RP values based on the two bootstrap methods changes when simulations are performed under the alternative hypothesis, resulting from changes in the observed test statistics with respect to the test threshold. Table 1 presents four samples close to the test threshold that reject and do not reject H_0 with sample sizes $n = 5$ and $n = 15$ for simulations under both the null and alternative hypotheses. This table includes the observed test statistics, test thresholds, PP-B-RP and NPI-B-RP. In the case of rejection, the PP-B-RP values tend to be higher than the NPI-B-RP values. Conversely, the values of PP-B-RP seem to be lower compared to the NPI-B-RP values in non-rejection cases. However, increasing n tends to reduce the differences between PP-B-RP and NPI-B-RP.

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

The two-sample t-test is commonly used to compare the means of two populations and is one of the most widely used statistical hypothesis tests. Known as the pooled variance t-test, it is applied when both samples meet the assumptions of normality, equal variances, and independence, as it is a parametric test [44]. Given two independent random samples, $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma^2)$, with unknown common variance σ^2, the hypotheses of interest are: $H_0 : \mu_1 = \mu_2$ against $H_a : \mu_1 \neq \mu_2$, $\mu_1 > \mu_2$, or $\mu_1 < \mu_2$, depending on the test direction.

Under the assumption of equal variances and normality, the test statistic is:

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

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

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

For a one-sided upper tail test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 > \mu_2$, we reject H_0 if $T > t_{n+m-2}^{(1-\alpha)}$, where $t_{n+m-2}^{(1-\alpha)}$ is the $(1-\alpha)$-th percentile of the t-distribution with $n+m-2$ degrees of freedom. For a one-sided lower tail test $H_a : \mu_1 < \mu_2$, we reject H_0 if $T < t_{n+m-2}^{(\alpha)}$. For the two-sided test $H_a : \mu_1 \neq \mu_2$, we reject H_0 if $|T| > t_{n+m-2}^{(1-\alpha/2)}$.

Welch introduced a version of the t-test for situations where the variances of two samples are significantly different [42]. Welch's t-test (also known as the unequal variance t-test) is suitable for comparing the means of two populations with unequal variances, assuming the samples are normally distributed. Let $X_1, X_2, \ldots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \ldots, Y_m \sim N(\mu_2, \sigma_2^2)$ be two independent samples from normal populations with unequal variances. The test statistic is:

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

where the degrees of freedom v are approximated by:

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

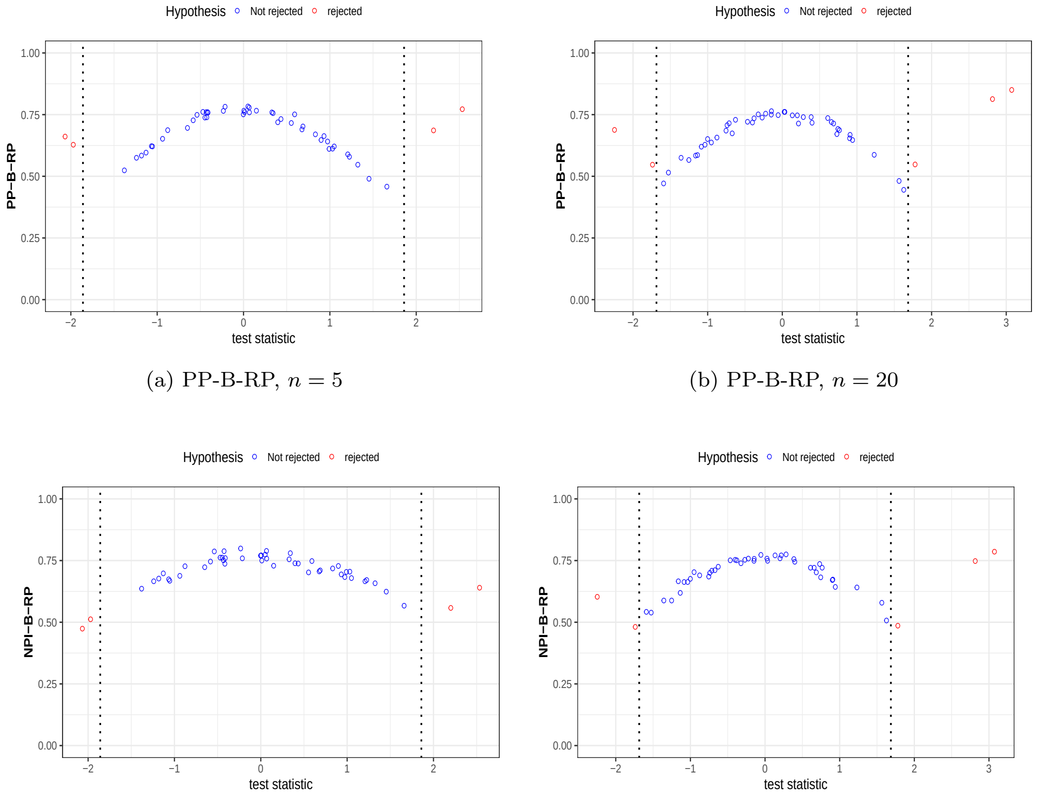
Unlike the Student's t-test, which assumes equal variances and estimates a pooled variance, Welch's t-test accounts for unequal variances. The degrees of freedom for Welch's test are typically smaller than those for the Student's t-test, making it more conservative [22].

The null hypothesis H_0 is rejected in favour of the one-sided upper tail test $H_a : \mu_1 > \mu_2$ at significance level α if $T > t_v^{(1-\alpha)}$, where $t_v^{(1-\alpha)}$ is the $(1-\alpha)$-th percentile of the Student's t-distribution with v degrees of freedom. For the one-sided lower tail test $H_a : \mu_1 < \mu_2$, reject H_0 if $T < t_v^{(\alpha)}$. For the two-sided test $H_a : \mu_1 \neq \mu_2$, reject H_0 if $|T| > t_v^{(1-\alpha/2)}$.

In this section, we examine the reproducibility (RP) of the two-sample t-test under the assumption of equal variances for both samples and compare it to Welch's t-test when the variances differ. While Student's t-test and Welch's t-test yield the same t-value, degrees of freedom, and p-value when sample sizes and variances are equal [21], differences in variances and/or sample sizes lead to variations in these metrics. The key distinction that led to the development of Welch's t-test is its accommodation of unequal variances and sample sizes. In such cases, the t-value remains the same, but the degrees of freedom and p-value differ. While Welch's t-test can be extended to more than two samples [43], we focus on the two-sample case with equal sample sizes. To evaluate the performance of the two bootstrap methods for RP in the two-sample t-test, we conduct simulation studies as follows:

1. Apply the t-test on the two original samples with equal sample sizes n, X and Y to obtain the value of the test statistic, then draw a conclusion about the null hypothesis for this test, whether it is rejected or not.
2. Draw a bootstrap sample of size n from sample X and a bootstrap sample of size n from sample Y. Apply the two-sample t-test to these two bootstrapped samples to obtain the test conclusion.
3. Perform Step 2 in total B times and record the test outcome each time whether or not the null hypothesis is rejected.
4. The ratio of B times that the two original samples and these two bootstrap samples have the same conclusion is the estimate of the RP.
5. Perform all these steps N times to obtain RP values for both rejection and non-rejection cases of the null hypothesis.

We first investigate the RP for the two-sample t-test when the variances of the two normally distributed populations are assumed to be equal. The



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

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

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

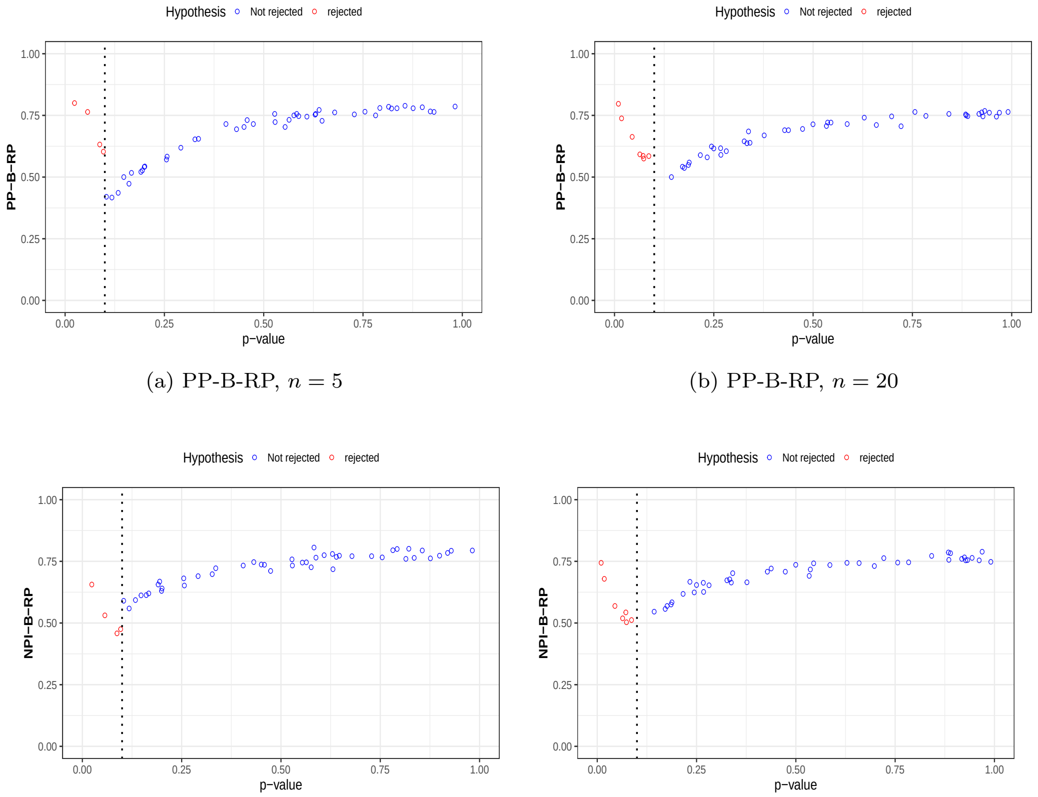
(d) NPI-B-RP, $n = 20$

Figure 3: Simulations under H_0: values of PP-B-RP and NPI-B-RP for two-sample t-test, where $n = 5, 20$.

two-sided two-sample t-test is considered, $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$, and level of significance $\alpha = 0.10$. We simulate two samples of size $n = 5$ under H_0 in total $N = 50$ times. The data are generated for the two original samples from the same Normal distribution with mean 2 and standard deviation 1. The RP value for the two-sample t-test is computed based on the two bootstrap methods as demonstrated above using $B = 1000$ bootstrap samples. The observed test statistic and Bootstrap-RP were determined for each of $N = 50$ samples. Also, we study the impact of increasing sample size to $n = 20$ on Bootstrap-RP values for the two-sample t-test. It is important to emphasise that the same data sets are used to compute the RP values for the two-sample t-test based on PP-B and NPI-B. The results of RP values based on the PP-B and NPI-B methods with samples of size $n = 5, 20$ under H_0 are presented in Figure 3.

The values of RP for both methods tend to increase as the test statistic moves away from the test thresholds, regardless of the decision on H_0. It is expected and rational, as discussed in Section 4. Increasing the size of samples leads to PP-B-RP and NPI-B-RP becoming close to 0.5 in both cases of rejection and non-rejection when the observed test statistics are close to the test thresholds. Also, the values of NPI-B-RP fluctuate narrowly as the sample size increases. These results happen with increasing the size of samples due to the decrease in the variability of the bootstrap samples and the increase in the power of the test. Simulation studies show that values of PP-B-RP have less variability than NPI-B-RP values, mainly when the sample size is small, due to the parametric model assumed for PP-B.

There is a tendency for PP-B-RP to be higher in cases of rejection than in non-rejection, whereas NPI-B-RP seems to be lower in cases of rejection than non-rejection. The reason for this is that the sample variance is included in the denominator of the test statistic for the two-sample t-test. The variance of PP-B is generally less than NPI-B due to the assumption of a parametric model in PP-B. For the upper tail test, PP-B samples lead to larger test statistic values than NPI-B samples due to a smaller variance value in the denominator. Therefore, the PP-B-RP tends to be lower in non-rejection cases due to the computed test statistic from PP-B samples tending to lie in the rejection region. Conversely, PP-B-RP tends to be higher in the case of rejection than NPI-B-RP because we obtain more cases that reject H_0. It is similar to what was discussed in Section 4 for the upper tail one-sample t-test. We can observe a similar impact on patterns of RP values based on PP-B and NPI-B for the lower tail test. It is important to note that the lower



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

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

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

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

Figure 4: Simulations under H_0: values of PP-B-RP and NPI-B-RP for Welch's t-test, where $n = 5, 20$.

tail two-sample t-test has negative values, which implies that PP-B samples lead to smaller test statistic values compared to NPI-B samples. Hence, we obtain a similar result to the upper tail test for PP-B-RP and NPI-B-RP.

Now, we consider the RP of the two-sample t-test when both samples are normally distributed with unequal variances. The procedure for determining the RP of Welch's t-test follows the same steps as for the two-sample t-test, except that we draw two original samples from Normal distributions with different standard deviations. Two samples of size n are simulated from two Normal distributions with different standard deviations, $\sigma_1 = 1$ and $\sigma_2 = 2$, but both with mean 2. A critical value of the test statistic for Welch's t-test is computed using the degrees of freedom which are random variables dependent on the size and variance of the sample. Therefore, we use the p-value for better visualization of figures rather than the critical value because each simulated sample has a different critical value even though all samples have the same size. The p-values and critical values are two different approaches that lead to the same result regarding whether the null hypothesis is rejected or not. Figure 4 shows the results of RP values for Welch's t-test using the two bootstrap methods with samples of size $n = 5, 20$ under H_0.

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

6. Bootstrap-RP for the F-test

The F-test for equality of variances tests the null hypothesis that the variances of two normal samples are equal. It is based on the ratio of the two sample variances, hence known as the F-ratio test. The F-test assumes normality for both samples and when this assumption is in doubt, alternative tests for variance comparison should be used [31]. Like the two-sample t-test and Welch's t-test, the F-test requires normality. Let $X_1, X_2, \dots, X_n \sim N(\mu_1, \sigma_1^2)$ and $Y_1, Y_2, \dots, Y_m \sim N(\mu_2, \sigma_2^2)$ be two independent random samples from normal populations. The hypotheses are:

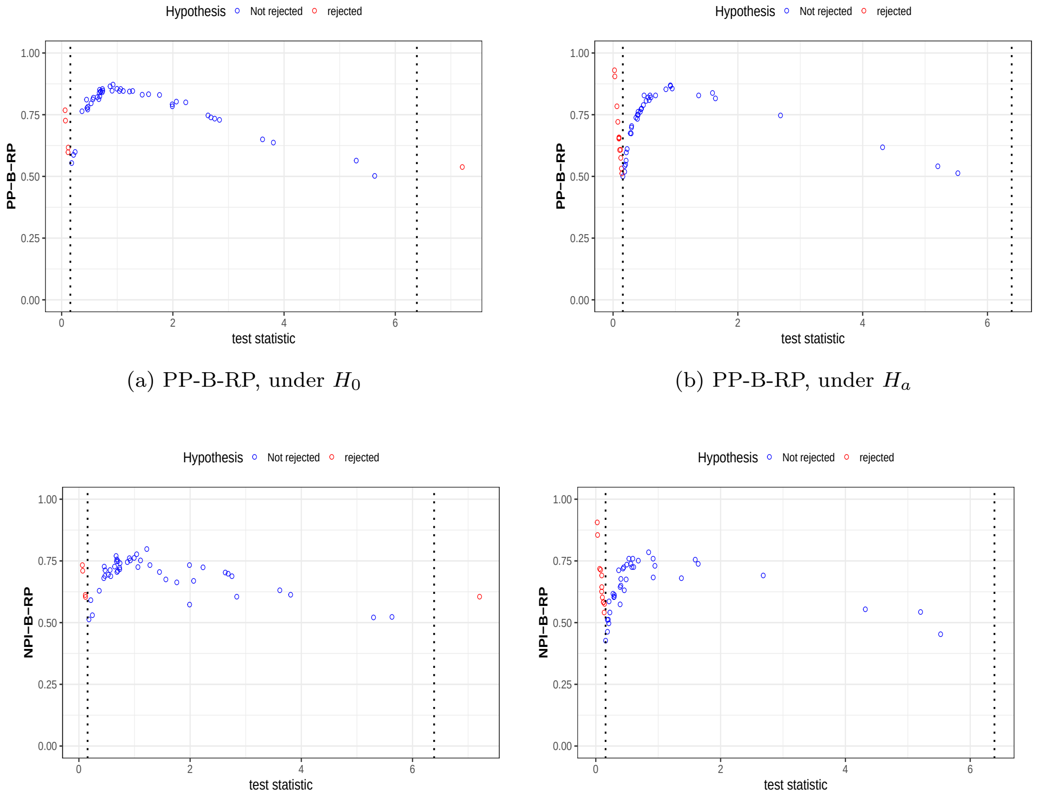
$$H_0 : \sigma_1^2 = \sigma_2^2 \quad \text{versus} \quad H_a : \sigma_1^2 \neq \sigma_2^2, \quad H_a : \sigma_1^2 > \sigma_2^2, \quad H_a : \sigma_1^2 < \sigma_2^2$$

The test statistic is $F = s_1^2/s_2^2 \sim F_{n-1,m-1}$, where $F_{n-1,m-1}$ is the F-distribution with $n-1$ and $m-1$ degrees of freedom. For a two-sided test, reject H_0 at significance level α if $F < F_{n-1,m-1}^{(\alpha/2)}$ or $F > F_{n-1,m-1}^{(1-\alpha/2)}$. For one-sided tests, reject H_0 if $F > F_{n-1,m-1}^{(1-\alpha)}$ for $H_a : \sigma_1^2 > \sigma_2^2$, and if $F < F_{n-1,m-1}^{(\alpha)}$ for $H_a : \sigma_1^2 < \sigma_2^2$.

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

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

A larger sample size is considered to study the effect of increased sample size on Bootstrap-RP values for the F-ratio test. Figure 6 presents the results



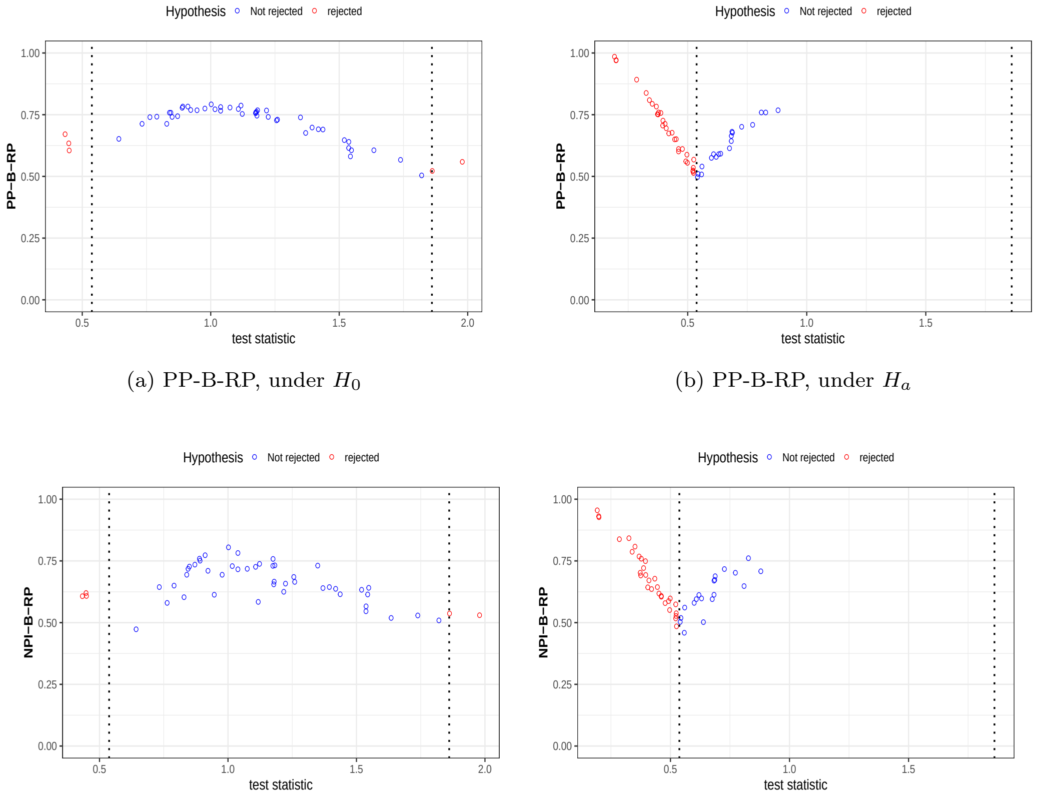
(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_a

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



(a) PP-B-RP, under H_0

(b) PP-B-RP, under H_a

(c) NPI-B-RP, under H_0

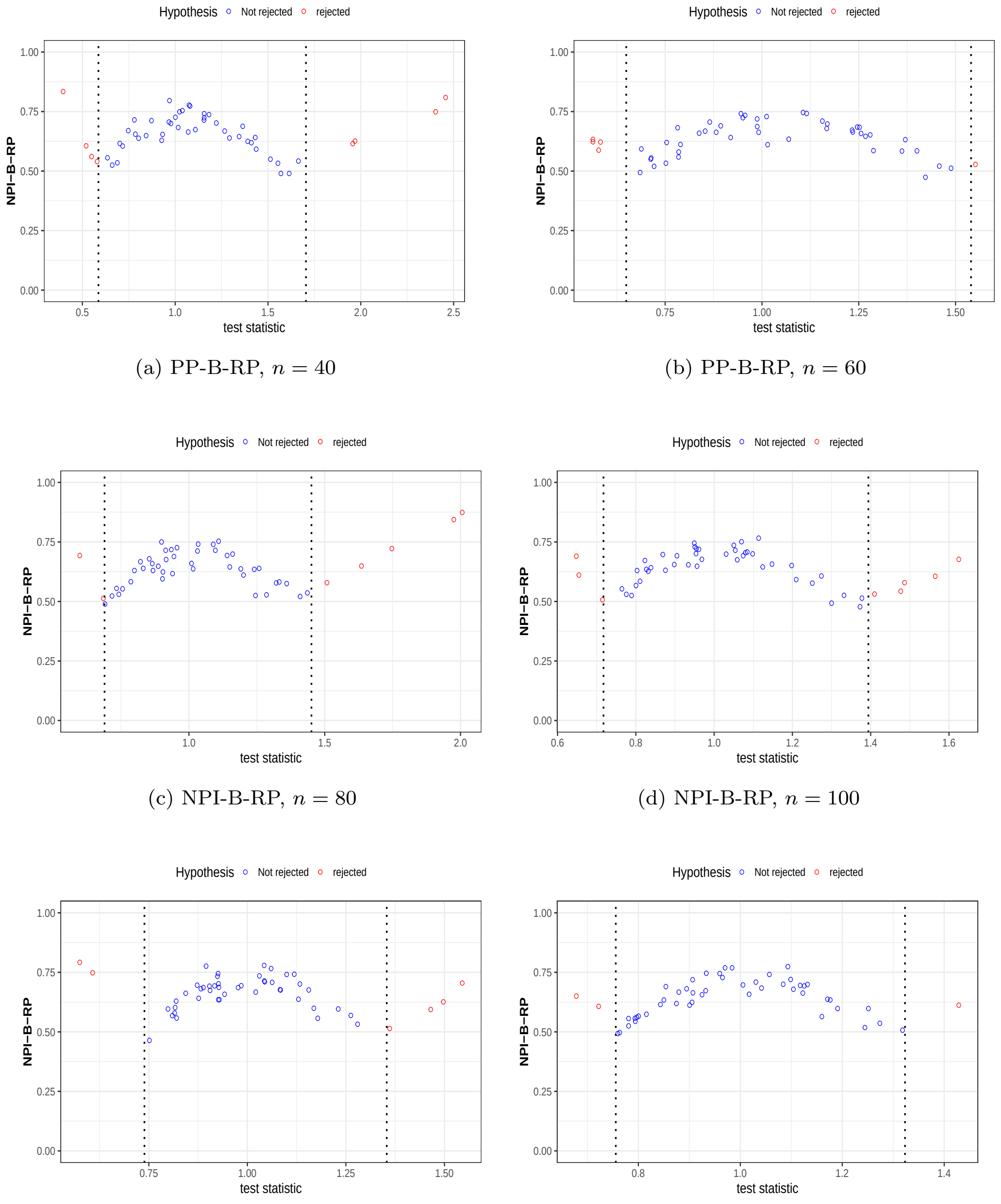
(d) NPI-B-RP, under H_a

Figure 6: Simulations under H_0 and H_a: values of PP-B-RP and NPI-B-RP for F-test, where $n = 30$.

of RP values using the PP-B and NPI-B methods for samples of size $n = 30$ under H_0 and H_a. As the size of the samples increases, the pattern of RP values changes under both the null and alternative hypotheses. We observe a change in the pattern of the RP values obtained through simulations under H_0 as the impact of the F-test follows F-distribution with larger degrees of freedom. Increasing the size of the samples leads to an increase in the power of the test, so we obtain more cases rejecting H_0 when simulations are performed by sampling under the alternative hypothesis. Simulations under H_a show changes in the pattern of the RP values due to changes in the observed test statistics in relation to the test threshold, as well as the effects of the F-test following the F-distribution with larger degrees of freedom. It is noteworthy that the variability of NPI-B-RP values is not reduced by increasing the size of the samples. Figure 7 presents additional results for the NPI-B-RP of the F-test, indicating substantial fluctuations even as sample sizes increase ($n = 40, 60, 80, 120, 140$). The NPI-B method exhibits greater variability than the PP-B method, as it does not rely on a parametric model. As a result, NPI-B-RP fluctuations for the F-test do not decrease with larger sample sizes, as the test statistic is merely the ratio of two sample variances.

7. NPI-RP and Bootstrap-RP for the likelihood ratio test

In this section, we study the RP of the likelihood ratio tests using the bootstrap method to compare it with the NPI-RP. The reproducibility probability of a test based on the NPI approach (NPI-RP) considers the test result for a predicted future sample of the same size as the original sample. This method is described in detail in Section 2. The exact NPI lower and upper reproducibility can only be computed for small data sets. Coolen and Marques [14] propose an alternative computational method to approximate NPI-RP for larger sample sizes via sampling of future orderings instead of considering all different possible orderings. They introduced sampling of orderings for the likelihood ratio test to overcome computational difficulties. In our work, we do not compute lower and upper reproducibility probabilities for the tests because it is hard to derive the minimum and maximum values of some test statistics, such as the test statistic of the t-test, which depend on both the sample mean and variance. However, we can construct the confidence interval for the single value of Bootstrap-RP using formula $\hat{p} \pm z^{(1-\alpha/2)}\sqrt{\hat{p}(1-\hat{p})/n}$, where the proportion $\hat{p}$ is the predictied Bootstrap-RP value. Here, we investigate whether or not the Bootstrap-RP tends to



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

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

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

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

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

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

Figure 7: Simulations under H_0: values of NPI-B-RP for F-test.

provide a value within the lower and upper NPI-RP.

Coolen and Marques [14] introduced sampling of future orderings for likelihood ratio tests with the test criterion in terms of the sample mean. The likelihood ratio test in the following test criterion involves the mean of the observed values. The null hypothesis H_0 is considered with a one-sided alternative hypothesis, $H_0 : \mu \leq \mu_0$ vs $H_a : \mu > \mu_0$, leading to the test criterion, H_0 being rejected if and only if

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

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

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

$$\underline{m}_j = \frac{1}{n}\sum_{i=1}^{n+1} s_i^j x_{(i-1)} \tag{2}$$

$$\overline{m}_j = \frac{1}{n}\sum_{i=1}^{n+1} s_i^j x_{(i)} \tag{3}$$

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

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\underline{m}_j > c\} \tag{4}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\overline{m}_j > c\} \tag{5}$$

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

The same arguments apply when the original data do not lead to the rejection of the H_0, allowing us to derive the NPI lower and upper probabilities for test reproducibility as

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\overline{m}_j \leq c\} \tag{6}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\underline{m}_j \leq c\} \tag{7}$$

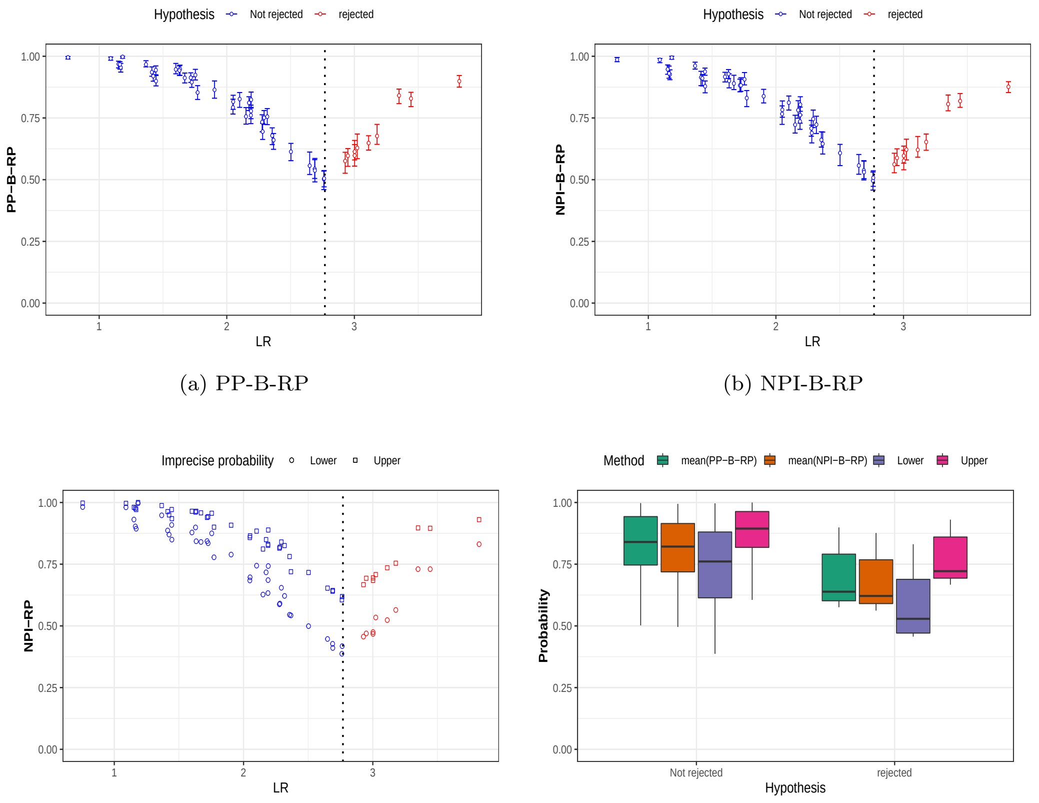
The decision rule may be expressed with the test criterion in terms of the sample mean $\overline{X}$ for the likelihood ratio test as test criterion (1), which rejects the null hypothesis for a significance level α if

$$\overline{X} > q_{(1-\alpha)} \tag{8}$$

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

We consider likelihood ratio tests for the mean value underlying the Normal population. For distributions with infinite range, we have to define bounds of possible values for the future observations, which we denote by $x_{(0)} = L$ and $x_{(n+1)} = R$. It is obvious that we must assume values $L < x_{(1)}$ and $x_{(n)} < R$ such that the observations are within this range $[L, R]$, where L and R can depend on the actual data observations. For n data observations $x_1 < x_2 < \ldots < x_n$, the lower and upper limits may be defined as $L = x_{(1)} - \frac{x_{(n)} - x_{(1)}}{n-1}$ and $R = x_{(n)} + \frac{x_{(n)} - x_{(1)}}{n-1}$.

We simulated $N = 50$ samples of size $n = 25$ from the Normal distribution with mean 2 and standard deviation 3 under H_0. We approximate NPI-RP for larger sample sizes via sampling of orderings instead of considering all different possible orderings. To achieve reasonable results, Coolen and Marques [14] suggest that the number of orderings sampled should be at least 2000. Considering the number of orderings sampled equal to 2000, the upper and lower RP for each of $N = 50$ samples were calculated based on the decision rule given in (8) with the level of significance $\alpha = 0.10$. The NPI lower and upper reproducibility probabilities are calculated for rejection cases using Equations (4) and (5). In the case of non-rejection, we compute



(a) PP-B-RP

(b) NPI-B-RP

(c) NPI-RP

(d) RP

Figure 8: Simulations under H_0: values of PP-B-RP, NPI-B-RP and NPI-RP for likelihood ratio test, where $n = 25$.

the NPI lower and upper reproducibility probabilities using Equations (6) and (7). We investigate whether or not the Bootstrap-RP methods tend to provide values that fall within the lower and upper NPI-RP for the likelihood ratio test. The RP for each of $N = 50$ samples was computed based on the PP-B and NPI-B methods using $B = 1000$ bootstrap samples. For each simulated sample, we compute RP values based on the bootstrap method and repeat the procedure 100 times, so we obtain $RP_1, \ldots, RP_{100}$. Then, we examine whether these values are between the corresponding lower and upper NPI-RP results. The same simulated samples are used to compute the RP values of the likelihood ratio test based on different bootstrap methods and NPI-RP. The observed likelihood ratio statistic, Bootstrap-RP, and NPI-RP were determined for each of the $N = 50$ samples.

Figure 8 presents RP values using different bootstrap methods and NPI-RP under H_0 for samples of size $n = 25$. The minimum, mean and maximum values of 100 Bootstrap-RP for each simulated sample are computed. The boxplots of RP are displayed for both rejections and non-rejections based on the mean of PP-B-RP and NPI-B-RP, as well as the lower and upper NPI-RP. We found 90% of PP-B-RP values and 88% of NPI-B-RP values are included in the bounds of NPI-RP. We conclude that both PP-B-RP and NPI-B-RP results are consistent with NPI-RP because most of these values are located in the corresponding NPI-RP boundaries. The PP-B-RP and NPI-B-RP are in line with NPI-RP in terms of investigating test reproducibility as a prediction problem rather than an estimation problem. Further simulations were performed under H_a, which led to similar results as the case presented under H_0.

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

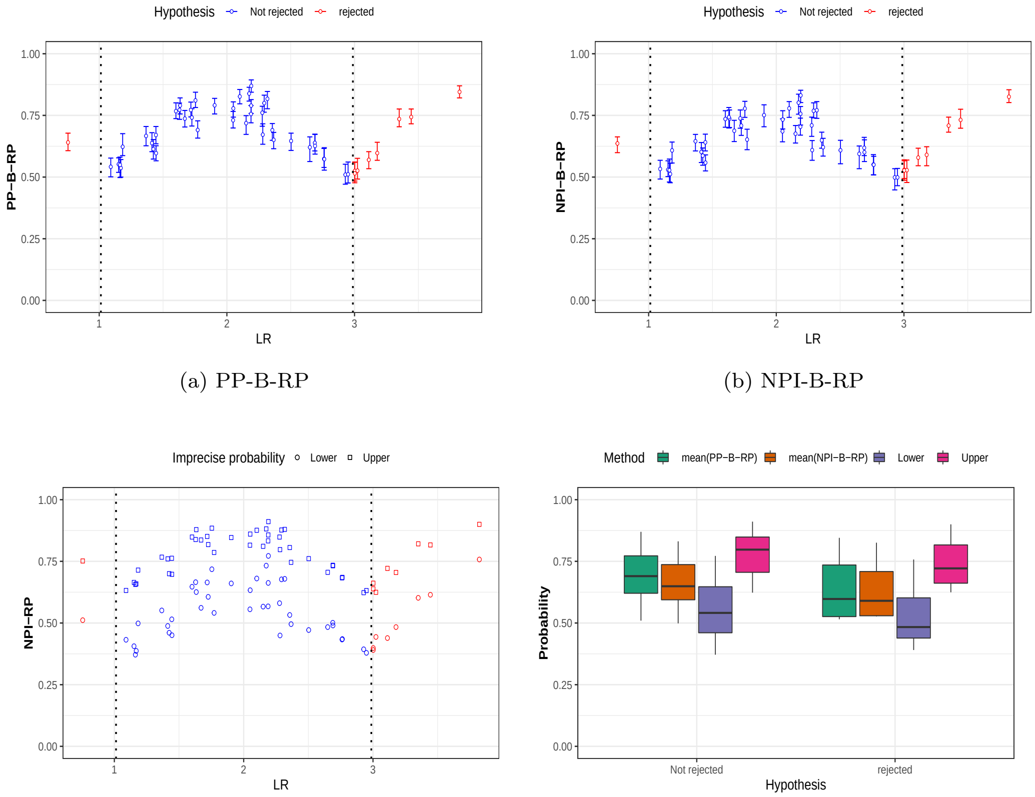
$$\overline{X} < q_{(\alpha/2)} \quad \vee \quad \overline{X} > q_{(1-\alpha/2)} \tag{9}$$

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

The minimum upper bound and maximum lower bound for the mean corresponding to O_j remain unchanged as in Equations (2) and (3), respectively. In the case of a two-sided test, the NPI lower and upper probabilities are different because they need to account for the two rejection regions. If the original data reject H_0, then the lower and upper RPs are derived as follows.

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\underline{m}_j > q_{(1-\alpha/2)} \quad \vee \quad \overline{m}_j < q_{(\alpha/2)}\} \tag{10}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\overline{m}_j > q_{(1-\alpha/2)} \quad \vee \quad \underline{m}_j < q_{(\alpha/2)}\} \tag{11}$$



(a) PP-B-RP

(b) NPI-B-RP

(c) NPI-RP

(d) RP

Figure 9: Simulations under H_0: values of PP-B-RP, NPI-B-RP, and NPI-RP for likelihood ratio test, where $n = 25$.

If the original data does not lead to rejecting the null hypothesis, we have

$$\underline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\underline{m}_j > q_{(\alpha/2)} \quad \wedge \quad \overline{m}_j < q_{(1-\alpha/2)}\} \tag{12}$$

$$\overline{RP} = \binom{n+m}{n}^{-1} \sum_{\boldsymbol{j}} \mathbf{1}\{\overline{m}_j > q_{(\alpha/2)} \quad \wedge \quad \underline{m}_j < q_{(1-\alpha/2)}\} \tag{13}$$

We have simulated $N = 50$ samples of size $n = 25$ from the Normal distribution with mean 2 and standard deviation 3 under H_0. For each case, we compute the lower and upper RPs for the two-sided test based on the decision rule given in (9) with the significance level $\alpha = 0.10$ by considering the number of orderings sampled equal to 2000. The lower and upper reproducibility probabilities of the NPI are computed for rejection cases using Equations (10) and (11). In the case of non-rejection, we calculate the NPI lower and upper reproducibility probabilities based on Equations (12) and (13). The same simulated samples are used to compute the RP values based on the bootstrap and NPI methods. We compute RP values for the two-sided test based on the bootstrap method and repeat the procedure 100 times for each simulated sample as we did with the one-sided test. Figure 9 shows RP values for the likelihood ratio test with the two-sided alternative using different bootstrap methods and NPI-RP under H_0 for samples of size $n = 25$. For each simulated sample, the minimum, mean, and maximum Bootstrap-RP values are computed. The boxplots of RP are shown in both cases of rejection and non-rejection based on the mean of PP-B-RP and NPI-B-RP, along with the lower and upper NPI-RP. All values of PP-B-RP and NPI-B-RP are included in the bounds of NPI-RP, indicating that these bootstrap methods align with the reproducibility probability based on the NPI approach. Further simulations were performed under H_a, yielding results similar to those of the case presented under H_0.

8. Conclusions and future works

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

The use of the bootstrap to predict RP avoids the hard calculations of the lower and upper boundaries in NPI-RP, and it offers a flexible approach when considering large sample sizes. The Bootstrap-RP uses the point estimate to present the RP instead of the lower and upper values of NPI-RP, but we can construct the confidence interval for the single value of Bootstrap-RP. We explore whether the RP values using PP-B and NPI-B tend to be between the lower and upper NPI-RP for the likelihood ratio test. The predicted values of PP-B-RP and NPI-B-RP for the likelihood ratio test are mostly included within the bounds of NPI-RP, indicating that these bootstrap methods are consistent with the NPI-RP approach. The PP-B-RP, NPI-B-RP, and NPI-RP consider test reproducibility from a predictive standpoint, which provides an appropriate formulation for inferring the RP of a test. It seems logical and natural to study the RP of a test with the same sample sizes and significance level as in the actual test. Senn [38] discussed how circumstances in the real world may vary among different tests, including sample sizes. The bootstrap method for the reproducibility of tests can be extended to consider future sample sizes that differ from the data sample size or to use varying levels of statistical significance. However, employing the same sample sizes and significance levels as in the actual test is logical from the perspective of theoretical reproducibility, particularly within a frequentist statistical framework.

Acknowledgements

This research took place during Abdulrahman Aldawsari's PhD studies at the Department of Mathematical Sciences, Durham University. Abdulrahman is grateful to Prince Sattam bin Abdulaziz University in Saudi Arabia for financial support (project number PSAU/2024/R/1446)

References

[1] A. S. M. Al Luhayb, F. P. A. Coolen, and T. Coolen-Maturi. Generalizing banks' smoothed bootstrap method for right-censored data. In *Proceedings of the 29th European Safety and Reliability Conference (ESREL 2019)*, pages 894–901, Hannover, Germany, 2019. ESREL 2019.

[2] A. Aldawsari. *Parametric Predictive Bootstrap and Test Reproducibility.* Doctoral thesis, Durham University, 2023.

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

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

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

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

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

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

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

[10] F. P. A. Coolen. Nonparametric predictive inference. In M. Lovric, editor, *International Encyclopedia of Statistical Sciences*, pages 968–970. Springer, Berlin, 2011.

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

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

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

[14] F. P. A. Coolen and F. J. Marques. Nonparametric predictive inference for test reproducibility by sampling future data orderings. *Journal of Statistical Theory and Practice*, 14(4):1–22, 2020.

[15] F. P. A. Coolen, M. C. M. Troffaes, and T. Augustin. Imprecise probability. In M. Lovric, editor, *International Encyclopedia of Statistical Science*, pages 645–648. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[16] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application.* Cambridge University Press, 1997.

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

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

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

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

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

[22] B. Derrick, D. Toher, and P. White. Why welch's test is type i error robust. *The Quantitative Methods in Psychology*, 12(1), 2016.

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

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

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

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

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

[28] P. Hall. Theoretical comparison of bootstrap confidence intervals. *The Annals of Statistics*, pages 927–953, 1988.

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

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

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

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

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

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

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

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

[37] M. L. Rizzo. *Statistical Computing with R.* Boca Raton ; London : CRC Press, 2008.

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

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

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

[41] A. Simkus, T. Coolen-Maturi, F. P. A. Coolen, and C. Bendtsen. Statistical perspectives on reproducibility: Definitions and challenges. *Journal of Statistical Theory and Practice*, to appear:Page Numbers, 2025.

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

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

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