The Impact of Preliminary Tests on Statistical Reproducibility of Location Tests

ALSHAHRANI, NORAH,DEREE,M

The Impact of Preliminary Tests on Statistical Reproducibility of Location Tests

Norah D. Alshahrani

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
University of Durham
England

July 2024

Dedication

This thesis is dedicated to my mother and my son Feras

The Impact of Preliminary Tests on Statistical Reproducibility of Location Tests

Norah D. Alshahrani

Submitted for the degree of Doctor of Philosophy
July 2024

Abstract

This thesis investigates statistical reproducibility (RP) as a predictive inference problem within the framework of nonparametric predictive inference (NPI). NPI is focused on the prediction of future observations using existing data. In this thesis, statistical reproducibility is defined as the probability of the event that, if the test was repeated under the same conditions and with the same sample size, the same test outcome would be obtained. This thesis presents contributions to NPI reproducibility for location tests and preliminary tests which are preliminary statistical analyses performed before the main or location hypothesis testing to evaluate assumptions for their validity. There is an ongoing debate about whether preliminary tests are necessary to validate assumptions for location tests; some argue they are important for optimal performance while others caution against their use.

This thesis aims to evaluate the RP for location tests, both with and without preliminary tests, aiming to examine the impact of preliminary tests on the RP for location tests. The potential impact of preliminary tests on RP of location tests is explored through simulation studies that compare RP of location tests with and without such preliminary tests.

The findings suggest that the impact of preliminary tests on RP for location tests is small, they do not substantially lead to improved or deteriorated RP of location tests.

Declaration

I, Norah Alshahrani, declare that 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. The work presented in this thesis is entirely my own, except where otherwise indicated.

Acknowledgements

I would like to extend my sincere thanks to everyone who played a role in completing this thesis. Their help, motivation and knowledge have been invaluable during this academic pursuit.

I am truly grateful to my dissertation advisors, Prof. Frank Coolen and Dr. Tahani Coolen-Maturi, for their guidance, mentorship, and unwavering support. Their expertise and insightful feedback have been instrumental in shaping this research.

I would like to acknowledge my country Saudi Arabia for their financial support, which made this research possible. Their commitment to advancing academic endeavours is commendable.

Many thanks to my family for their encouragement and support during this period of my studies. Their support has been a motivation for progress and perseverance

Special appreciation to my colleagues for their friendship, discussions and shared ideas. Their collaborative spirit was a source of inspiration and motivation.

Finally, I thank everyone who contributed to this academic endeavour. Your contributions have left an indelible mark on this journey.

Contents

Chapter 1

Introduction

1.1 Motivation

One of the forms of statistical inference is hypothesis testing, where researchers in hypothesis testing create different hypotheses about a population parameter and then analyze sample data to determine which hypothesis is more supported by the evidence. This procedure includes establishing a null hypothesis, which usually represents a default or no-effect situation, and an alternative hypothesis, which reflects the researcher's speculation or the existence of an effect [97]. In applied research, it is common practice to conduct data analyses using multi-stage procedures of hypothesis testing. These procedures involve the implementation of one or more preliminary tests before carrying out the main tests of interest [53, 56]. This thesis restricts attention to location tests as the main tests and preliminary tests of Normality and equality of variances. The purpose of these preliminary tests is to evaluate certain assumptions or conditions necessary for the location tests. One prevalent example of the multi-stage procedure is the comparison of two population means using Student's t-test. In this case, researchers often begin by examining the Normality assumption through a preliminary goodness-of-fit test. If the null hypothesis of Normality is rejected, then an alternative nonparametric test, such as the Mann-Whitney test, is employed to analyze the data. On the other hand, if the null hypothesis of Normality is not rejected, then further tests may be conducted to assess additional assumptions, such as homogeneity of variances. Based on the outcomes of these preliminary tests, researchers make decisions regarding the appropriate statistical analysis to be performed [53].

Many researchers have discussed the importance of applying preliminary tests such as Normality and equality of variances tests to ensure the appropriate application of statistical methods. For example, Micceri's study [69] investigative achievement and psychometric measures revealed significant non-Normality in all 440 measures, including bimodality, exponential level asymmetry, and tail weights ranging from the uniform to the double exponential. Current studies have further demonstrated that the majority of real data samples show at least slight deviations from Normality in terms of skewness and kurtosis such as the study by Blanca et al[14]. Ruscio and Roche [83] also showed in their research the assumption of variance homogeneity is often violated in published studies.

However, authors such as Wells and Hinze [102] and Schucany and Ng [84] highlight several theoretical drawbacks against using preliminary testing. One essential issue is the implicit nature of conclusions drawn from preliminary tests rather than explicit ones. Acceptance of the null hypothesis, for example, the Normality assumption, based on insufficient evidence may lead to unwarranted assumptions about the data distribution. Additionally, assumptions underlying preliminary tests themselves raise questions about their validity, thus creating a paradoxical situation. In practice, in small to moderate sample sizes, preliminary tests may not guarantee alignment between sample and population characteristics. Altman [4] emphasizes how even samples drawn from theoretically Normal populations can show non-Normality. Furthermore, the application of preliminary tests on the same data as subsequent analyses introduces the risk of uncontrolled error rates, as emphasized by Schucany and Ng's simulation study [84]. More recent studies further question the necessity of preliminary testing. One of these is the study conducted by Shamsudheen and Hennig [86], their conclusions offer a broader consideration of whether preliminary tests should be used in applied statistics. They argue that while these tests can provide useful information, they may not always be necessary and can add unnecessary complexity and potential errors. Additionally, studies by Zimmerman [104, 105], Rasch, Kubinger, and Moder [78], and Rochon, Gondan, and Kieser [80] suggest that preliminary testing can not only be pointless but also result in inflated error rates and suboptimal test selections.

In light of these critiques, the motivation for this research is to assess the impact of preliminary tests on the reproducibility probability (RP) of location tests. reproducibility probability (RP) is the probability of the event that if the test was repeated under the same circumstances and with the same sample size, the same test outcome would be obtained. The reproducibility probability is a helpful indicator of the confidence of the results of statistical tests. Thus, it is essential to understand the potential impact of these preliminary tests on the RP of the location tests. This raises questions about the robustness of research findings when conducting preliminary tests before the location tests. Does the inclusion of preliminary tests affect the RP of the location tests? Are there differences between RP for location tests with preliminary tests and RP for location tests conducted alone, without preliminary tests?

In this thesis, the reproducibility is assessed from a nonparametric predictive inference (NPI) perspective. NPI is a frequentist approach that relies on a few assumptions and focuses on making predictions for future observations [21]. The predictive nature of NPI provides a natural formulation of inference on reproducibility which is an important characteristic of statistical test outcomes. We consider RP within a frequentist statistical framework from the perspective of prediction [24].

1.2 Location tests

Location tests are statistical tests used to determine if there is a significant difference between two or more populations or samples. They are often used in hypothesis testing to compare means, medians, or other measures of central tendency between groups. There are two different types of tests [19, 34]: parametric and nonparametric location tests. The main difference between them is based on the fundamental assumptions regarding the data under analysis [19, 34]. Parametric statistics deal with numerical data that follow continuous and known distributions. When the data are on an interval or ratio scale and the sample size is large, parametric statistical methods are suitable. However, when the data do not conform to a known distribution, nonparametric statistics, also known as distribution-free methods, become appropriate. Parametric approaches make assumptions about the sample population's underlying distribution, while nonparametric methods do not make any assumptions regarding the underlying population's distribution. Parametric tests assume that data follow a Normal distribution at the interval/ratio level of measurement. In contrast, nonparametric methods do not assume Normality or any specific distribution for the sample population and are often based on ranked or nominal data instead of actual measurements [19, 34]. In numerous instances, a parametric test can be replaced with a non-parametric test. Some examples of parametric tests are the z-test, the Student's t-tests, Analysis of Variance (ANOVA), and linear regression. Whereas the Mann-Whitney U test, the Kruskal-Wallis test, the chi-square test for association, the chi-square test for goodness-of-fit, and Spearman's rank correlation are some examples of nonparametric tests [19, 34].

The biggest advantage of parametric tests is that they are more powerful and precise than nonparametric tests, meaning they have a better chance of identifying a true effect or distinction in case it exists [99]. However, the biggest drawback of parametric tests is that they are sensitive to violations of the assumptions of Normality and homoscedasticity [66]. The biggest advantage of nonparametric tests is that they are more robust and flexible than parametric tests, where they can deal with data that are skewed and have outliers. Moreover, they also do not need large samples or random sampling to be valid[99]. However, the biggest drawback of nonparametric tests is that they are less precise and powerful than parametric tests [99].

1.3 Preliminary tests in statistical analysis

A preliminary test is conducted to verify assumptions or conditions necessary for the validity of subsequent statistical tests. Where the results of preliminary tests determine which statistical test should be applied to evaluate the main hypothesis [47, 67]. Boon [16] described the preliminary testing procedure as follows: If we need to test a hypothesis and we are unsure whether to use a restricted but potentially incorrect model or a larger and less precise model. In order to resolve this matter, we conduct a preliminary test on the suitability of the restricted model. If this test does not show significance, we can continue using the restricted model and perform a test designed for that model. Otherwise, we can use a different main test that is more suitable for the larger model but has less strength compared to the first test when the restricted model is valid.

Common types of preliminary tests before using parametric statistical tests are Normality tests and equality of variances tests. Normality tests help to assess whether data follow a Normal distribution and are usually a frequent preliminary step. Tests like the Shapiro-Wilk test [87] or Kolmogorov-Smirnov test [58] are commonly used for this purpose. Equality of variances tests are another important preliminary test for parametric tests. Tests like Levene's test [61] or F-test [93] help determine if there is a significant difference between the variances across groups.

Applying a preliminary test for Normality to choose between a parametric test and a nonparametric test is referred to as the two-stage procedure by some researchers. For example, the two-stage procedure involving a preliminary Normality test is common for choosing between the two-sample t-test and the Mann-Whitney U test for independent samples [40, 80]. Freidlin et al. [41] referred to this as a "natural adaptive procedure" where the Shapiro-Wilk for the Normality test is applied, and if the null hypothesis of Normality is not rejected, the t-test is used. Otherwise, the non-parametric Wilcoxon rank-sum test is applied. Similarly, performing preliminary tests for both Normality and equality of variances to choose between a parametric test and a nonparametric test is referred to as the three-stage procedure [53, 78, 80].

1.4 NPI-RP

This section presents the basic concept of nonparametric predictive inference (NPI) as the main approach used in this thesis. In addition, the NPI bootstrap method (NPI-B) and statistical reproducibility (RP) are presented, specifically focusing on NPI bootstrap reproducibility probabilities (NPI-B-RP).

1.4.1 Nonparametric Predictive Inference (NPI)

Nonparametric predictive inference (NPI) is a frequentist statistical method that relies on Hill's assumption $A_{(n)}$ [52]. This assumption provides a direct probability for future observations of a random quantity, conditional on the observed values of related random quantities [21]. Let n real-valued ordered observations $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, where $n \geq 1$, correspond to continuous and exchangeable random quantities $X_1, \ldots, X_n, X_{n+1}$ enumerated in increasing order. For convenience, let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$. In the case of non-negative random quantities, we set $x_{(0)} = 0$. It is important to note that $x_{(n+1)}$ does not represent an observed value for X_{n+1}. These n observations divide the real line into $n+1$ intervals $I_j = (x_{(j-1)}, x_{(j)})$, where $j = 1, 2, \ldots, n+1$. The assumption $A_{(n)}$ regarding a future observation X_{n+1} based on n observations can be expressed as:

$$P(X_{n+1} \in I_j = (x_{(j-1)}, x_{(j)})) = \frac{1}{n+1}, \quad \text{for } j = 1, 2, \ldots, n+1. \tag{1.4.1}$$

This implies that X_{n+1} has an equal chance of being within any of the intervals created by the ordered observed data. $A_{(n)}$ does not make any additional assumptions and can be interpreted as a post-data assumption regarding exchangeability [21]. It assumes no ties in the data or predictions. In the case of tied observations, it can be dealt with by breaking them by a tiny amount [28, 29]. If we want to allow ties, the probabilities $\frac{1}{n+1}$ can be assigned to closed intervals $I_j = [x_{(j-1)}, x_{(j)}]$ instead of open intervals $I_j = (x_{(j-1)}, x_{(j)})$ [20].

The NPI approach can be extended to incorporate $m > 1$ future observations by consecutively applying Hill's assumption $A_{(n)}, A_{(n+1)}, \ldots, A_{(n+m-1)}$, denoted as $A_{(.)}$ [30]. Let O_i represent the possible orderings of the $m > 1$ future observations relative to the existing n data observations. There are $\binom{n+m}{n}$ possible orderings O_i, where $i = 1, 2, \ldots, \binom{n+m}{n}$, and under $A_{(.)}$, each ordering is equally likely [30]. For a specific ordering O_i, let S_j^i denote the number of future observations in the interval $I_j = (x_{(j-1)}, x_{(j)})$, where $j = 1, 2, \ldots, n+1$. The variable S_j^i quantifies the count of future observations that fall within each interval defined by the existing data. Then, based on the $A_{(n)}$ assumptions, we have:

$$P\left(\bigcap_{j=1}^{n+1} S_j^i = s_j^i\right) = P(O_i) = \frac{1}{\binom{n+m}{n}}, \qquad i = 1, \ldots, \binom{n+m}{n} \tag{1.4.2}$$

where s_j^i are non-negative integers and $\sum_{j=1}^{n+1} s_j^i = m$. Specific ordering only indicates the number of future observations in each interval I_j, without making any assumptions about the exact location of the future observations within the interval I_j.

In the NPI framework, uncertainty is typically represented using lower and upper probabilities [8]. This approach does not focus on the exact positioning of future points.

Instead, it emphasizes that a future point falls within an interval I_j delimited by two consecutive observations $x_{(j-1)}$ and $x_{(j)}$. The lower probability of event A represents the maximum lower bound for its precise probability. In statistical hypothesis testing, this event could involve either rejecting or not rejecting the null hypothesis. In simpler terms, a lower reproducibility probability indicates strong evidence in support of event A [8]. Within the NPI framework, lower probability considers only how m future observations are ordered among the n current observations where event A must occur [22]. The upper probability represents the minimum upper bound for event A, taking into consideration all potential evidence favouring event A. Within the NPI framework, the upper probability considers all possible orderings in which event A could occur [22].

Exchangeability in the NPI framework

Hill's assumption requires that random variables are exchangeable. Exchangeability does not necessarily indicate a type of dependence, as it is important to be able to learn from observations about unobserved random variables without assuming any specific form of dependence [90]. Additionally, exchangeability does not mean independence. For instance, if random variables X and Y are independent, then acquiring information about X would not affect our knowledge or beliefs regarding Y. The concept of exchangeability is utilized in situations with limited knowledge about the relevant random variable or when a deliberate decision has been made to exclude this information, making independence an unsuitable assumption [90].

In the context of the NPI framework, exchangeability suggests that, for real-valued quantities, the orderings are equally probable before observing their values [90]. In a frequentist statistics scenario, $A_{(n)}$ fills in the values of n observations and thus results in a $\frac{1}{(n+1)}$ probability for a future observation to fall within each interval between two consecutive observations. If one were to propose a minimal formulation, $X_1, \ldots, X_{(n+n)}$, for n future observations, would not necessarily need to be exchangeable as only the assumptions related to $A_{(.)}$ are required. Therefore, it may not be essential for the first n observed quantities' exchangeability. However, assuming all random quantities to be exchangeable still holds logical significance [90].

1.4.2 NPI-Bootstrap (NPI-B)

Nonparametric predictive inference bootstrap (NPI-B), introduced by Coolen and Bin-Himd [24], is a distinctive bootstrap method designed specifically for predicting future observations within the framework of nonparametric predictive inference (NPI). NPI-B differs from other bootstrap methods that aim to estimate the characteristics of an assumed underlying population distribution, as NPI-B explicitly quantify the uncertainty in predicting future values [24].

NPI is a frequentist statistical framework explicitly focused on predicting future observations, enabling the development of a bootstrap method tailored for this purpose (NPI-B) [24]. The NPI-B method relies on Hill's assumption $A_{(.)}$, which states that a future observation is equally likely to fall into any of the $n + 1$ intervals created by the n data points [25]. Notably, NPI-B does not require assuming an underlying distribution for the data, making it a completely nonparametric method [25]. NPI-B samples are not drawn from the original data sample itself but from the predictive distribution for future values given the data [25].

The NPI-B approach involves n data observations and focuses on predicting m future observations. Let the number of bootstrap samples be denoted as N. The NPI-B approach for real-valued data on a finite and an infinite interval works as follows: [24, 25].

1. Take n ordered observations $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, assuming there are no ties

2. These n observation create $n + 1$ intervals.

3. Choose one interval of $n + 1$ intervals randomly.

4. If this interval is of finite length, sample an observation Uniformly. If the interval is unbounded, such as $(-\infty, x_{(1)})$ or $(x_{(n)}, +\infty)$, sampling an observation involves assuming the tails of a Normal distribution over these intervals. While if data is on the non-negative real line (infinite interval $[0, +\infty)$) and the interval is $(x_{(n)}, +\infty)$, we sample the future value from the tail of Exponential distribution.

5. Add this observation to the data n, leading to $n + 1$ observations in the data set.

6. Perform Steps 2-5, now with order $n + 1$ data, to obtain a further future value.

7. Following Steps 2-6, in total m times to form an NPI-B sample.

8. Create more NPI-B samples of size m N times.

In Step 4, since it cannot be sampled observation Uniformly from an open-ended interval $(-\infty, x_{(1)})$ and $(x_{(n)}, \infty)$, it is assumed to sample the future observation over these intervals from the tails of a Normal distribution, with an estimated mean (μ) which is calculated as the midpoint of the interval $\mu = \frac{x_{(1)}+x_{(n)}}{2}$, and estimated standard deviation (σ) is computed as $\frac{x_{(n)}-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ represents the CDF of the standard Normal distribution [25]. σ is estimated using the properties of the Normal cumulative function: $P(Y > x_{(n)}) = 1 - \Phi(\frac{x_{(n)}-\mu}{\sigma}) = \frac{1}{1+n}$ [25]. For non-negative real-valued data, the NPI-B algorithm uses the tail of an Exponential distribution with the estimated rate $\lambda = \frac{\ln(n+1)}{x_{(n)}}$. λ is estimated using $P(Y < y) = 1 - e^{(-\lambda y)}$, given $P(Y > x_{(n)}) = \frac{1}{n+1}$ [25].

1.4.3 Statistical reproducibility

Reproducibility denotes the ability of future an experiment or studies to replicate the same results as the original experiment result. It plays an important role in scientific methods, providing researchers with confidence in the validity of their findings. Recently, reproducibility has gained increased attention within the scientific community, sparking extensive discussions on its various aspects. Atmanspacher and Maasen [7] offers an overview of many such facets. While the focus has largely been on topics like publication bias and guidelines for best practices to prevent major reproducibility challenges, surprisingly little attention has been given to the reproducibility of statistical inference methods—often a key component in investigations [26].

Goodman [44] was the first to discuss the concept of reproducibility probability within a hypothesis testing framework. He addressed a potential misunderstanding about the interpretation of statistical p-values, specifically that a low p-value does not necessarily improve the credibility of the test result. Goodman argued that the replication probability may be smaller than expected, highlighting the need for careful interpretation of p-values to avoid misinterpretation. Goodman used the term replication probability instead of reproducibility probability (RP). Goodman [44] defined statistical reproducibility as the probability of observing a similarly significant outcome in the same direction if the experiment was to be repeated under the same circumstances and with the same sample size. In a later extensive discussion of Goodman's paper, Senn [85] emphasizes the difference between p-values and reproducibility probability (RP). While Goodman argues that p-values can overstate the evidence against the null hypothesis, Senn argues that the issues with p-values primarily arise from their misinterpretation and misuse rather than the metric itself. Senn stresses the importance of distinguishing between the evidence provided by p-values and the concept of reproducibility.

De Martini [38] assumed the power of the test as an estimator of RP to evaluate results across a wide range of parametric tests. Moreover, he suggested defining statistical tests themselves using the estimated RP. De Capitani and De Martini further explored this power-based approach [35, 36, 37], who extended it to the estimation of RP for various nonparametric tests, including the Wilcoxon signed-rank test, sign test, Kendall test, and binomial test. In this approach, the power of a test is defined as the probability of rejecting the null hypothesis when the alternative hypothesis is true. By leveraging the power of the test as an estimator, researchers could infer the likelihood of reproducibility based on the test's ability to detect true effects.

Recently, a new perspective on RP was introduced by Coolen and Binhimd [24], employing the nonparametric predictive inference (NPI) framework of frequentist statistical methods. With its explicitly predictive nature, the NPI framework provides a natural way to make inferences about RP. By leveraging the predictive nature of NPI, it becomes possible to predict the outcome of a future test based on the data from the initial test, assuming the future test is conducted under the same conditions and with the same sample size as the initial test [13]. Coolen and BinHimd [24, 25] pioneered the use of NPI in testing reproducibility, including investigating NPI reproducibility for basic nonparametric tests like the Wilcoxon Mann–Whitney test (WMT) and developing NPI bootstrap for predictive inference [25]. Moreover, Alqifari and Coolen [3, 23] extended NPI reproducibility to testing on population quantiles as well as for a precedence test. Moreover, Simkus et al. [91] investigated NPI reproducibility in making final decisions based on the outcomes of multiple t-tests, highlighting the potential challenge of reproducibility in multiple testing scenarios. This thesis adapts Coolen and Binhimd [24] approach to statistical reproducibility.

In the context of Bayesian analysis, Billheimer [12] has examined predictive inference. To enhance reproducibility, he recommended using predictive inference to predict future observable data rather than infer unobservable parameters. This approach aligns with the approach to statistical reproducibility proposed in this thesis; however, it suggests utilizing the NPI framework instead of the Bayesian framework due to its not making assumptions about the data.

1.4.4 NPI-RP

Senn [85] argued that the probability of reproducibility of a hypothesis test could be as low as 0.5 in the most unfavourable case, particularly when a test statistic is near the threshold value between rejecting and non-rejecting a null hypothesis. Coolen and Bin Himd [24] confirmed this for certain fundamental tests related to one group of data or population, finding that with minimum NPI lower reproducibility probability was 0.5 [26]. Whereas when conducting basic tests with two groups of data or populations, the minimum NPI lower reproducibility probability was below 0.5, and the reproducibility tended to be worse if the null hypothesis was rejected and the test statistic was close to the threshold. This issue is compounded by the design of hypothesis tests, which are often tailored towards rejecting the null hypothesis, aligning with a key aim of many experiments [26]. Additionally, there is concern that both NPI lower and upper reproducibility probabilities can remain relatively low for test statistic values far from their respective thresholds[26].

The NPI-based method for assessing the reproducibility of statistical hypothesis tests involves conducting the test on the original data and then considering its results for all potential future data sets of similar size, based on the assumption of post-data exchangeability [26]. However, this approach presents computational difficulties for complex tests or larger data sets [26]. If it is possible to determine whether an ordering of future observations among the original data will result in rejection or non-rejection of the null hypothesis without assuming specific values between two original observations, sampling of the future orderings offers a solution leading to estimates of lower and upper reproducibility probabilities according to NPI [26, 27]. In cases where conclusions of the hypothesis test about a future data set can only be drawn from the precise knowledge of the future observations, applying the NPI bootstrap method [13, 25] is recommended [26].

1.4.5 NPI-B-RP

This thesis uses the nonparametric predictive inference bootstrap to estimate reproducibility probability (NPI-B-RP). NPI-B present a point estimate of reproducibility probability, it does not enable providing of the results in terms of imprecise reproducibility probabilities. The NPI-B-RP method takes advantage of the NPI-B framework for prediction, allowing the estimation of RP by reiterating a statistical test multiple times and observing the consistency of outcomes [13].

BinHimd [13] conducted a brief study on NPI-B-RP for the Wilcoxon Mann–Whitney (WMT) test to demonstrate that NPI-B-RP yields results in line with theoretical values for both lower and upper values of NPI-RP that can be calculated using small sample sizes. Moreover, Simkus [90] adopted the BinHimd algorithm for the NPI-B-RP for the Wilcoxon Mann-Whitney to calculate NPI-B-RP for the t-test with some differences such as the number of times the entire bootstrap process is repeated h and report NPI-B-RP using various statistics (minimum, mean, maximum).

In this thesis, it is difficult to calculate precise values of lower and upper reproducibility probabilities for the two-stage procedure testing or three-stage testing (location tests with preliminary tests). Therefore, we depend on the NPI bootstrap to estimate and assess the reproducibility. Algorithm 1 for calculating NPI-B-RP for the location tests without preliminary tests and preliminary tests only have been derived from the NPI-B-RP for the WMT test, which was outlined in BinHimd's thesis [13] and from the NPI-B-RP for t-test, which was present in Simkus's thesis [90]. We will rely on Simkus [90] mothed to report NPI-B-RP. Thus, Algorithm 1 is applied with $N = 1000$ and $h = 100$ (these numbers are optional and can be changed), and min, mean, and max of $\mathrm{RP}_1, \mathrm{RP}_2, \ldots, \mathrm{RP}_h$ were chosen as outputs.

Simkus [90] discussed in her thesis the reasoning behind the presentation of algorithm outputs and the selection of values for N and h. Various summary statistics were investigated through simple simulations, including the minimum, mean, median, maximum, as well as the 5th and 95th percentiles (representing the bootstrapped 90% confidence in-

Algorithm 1: NPI-B-RP for an interest test

```
Require: original samples, $N$, and $h$.
1: Apply a test on the original samples, and make a decision about $H_0$, $T$ symbolizes the decision, then record $T = 1$ if $H_0$ is rejected at significance level $\alpha$ and otherwise record $T = 0$.
2: Draw an NPI-B sample $N$ times based on the original samples, with sample size as original samples, and apply the same test. Each time record the test decision $T_j$, where $j = 1, 2, \ldots, N$, where we record $T_j = 1$ if $H_0$ is rejected or $T_j = 0$ if $H_0$ is not rejected.
3: compute RP, where $RP = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}_{(T=T_j)}$.
4: Perform steps 2-3 in total $h$ times, leading to $RP$ values $RP_1, \ldots, RP_h$.
```

terval) of $\mathrm{RP}_1, \mathrm{RP}_2, \ldots, \mathrm{RP}_h$. Simkus [90] found that there was no substantial benefit in using the 90% confidence interval with the maximum and minimum values, as the difference between the minimum value and the 5th percentile, and between the 95th percentile and the maximum value was negligible. Similarly, reporting the median did not provide additional insight, as the mean and median values of $\mathrm{RP}_1, \mathrm{RP}_2, \ldots, \mathrm{RP}_h$ were very similar. Thus, the mean value was considered the most important indicator of NPI reproducibility, referred to as the NPI-B-RP value. Additionally, the minimum and maximum values of $\mathrm{RP}_1, \mathrm{RP}_2, \ldots, \mathrm{RP}_h$ were reported alongside the NPI-B-RP value. Simkus [90] chose $N = 1000$ and $h = 100$, the primary objective was to strike a balance between computation time and accuracy. The value of h is set as 100, increase the value of h from 100 to 200 or even up to 500 leads to widening the gap between the minimum and maximum values of $\mathrm{RP}_1, \mathrm{RP}_2, \ldots, \mathrm{RP}_h$; however, this alteration is minimal and only causes a slight difference in the mean value at the third decimal place. Larger h leads to proportionally longer computational times without substantially enhancing accuracy. Simkus [90] found that when N was raised from 1000 to $10,000$, the means of $\mathrm{RP}_1, \mathrm{RP}_2, \ldots, \mathrm{RP}_h$ remained similar, differing only in the third decimal place, indicating that the algorithm performed well at $N = 1000$.

In this thesis, we have made several important contributions to the field of nonparametric predictive inference (NPI) and its application to statistical reproducibility, particularly in the context of preliminary and location tests. One primary contribution is the application of the NPI-B method, introduced by Coolen and BinHimd [24], to estimate the reproducibility of various preliminary tests, such as tests of Normality (Shapiro-Wilk test, Anderson-Darling test, and Lilliefors test) and equality of variances (F-test and Levene's test), as well as location tests (one-sample t-test, one-sample Wilcoxon test, Welch's t-test, ANOVA test, Welch's ANOVA test, and Kruskal-Wallis test). Another key contribution is the estimation of reproducibility probability for two-stage and three-stage testing procedures through the use of the NPI-B method. These multi-stage procedures are designed to sequentially apply preliminary tests and location tests, providing a structured framework for hypothesis testing that may enhance the reproducibility probability (RP) of the results.

1.5 Outline of thesis

This thesis is organised as follows: Chapter 2 investigates the reproducibility (RP) of Normality tests. We employ three well-known Normality tests which are Shapiro-Wilk (SW), Anderson-Darling (AD), and Lilliefors (LF) to examine reproducibility under different distributions and sample sizes and compare their RP performance. Additionally, RP for the Normality tests are explored under different levels of significance. The relationship between the overall mean of RP values in the rejection area and the estimated power for the Normality tests is examined. The reproducibility of the Normality tests is studied because it is considered an essential step in the topic of this thesis "the impact of preliminary tests on RP of location tests", where the location tests require investigation of the Normality assumption, investigated through Normality tests.

In Chapter 3, the reproducibility of equality of variances tests is investigated. The primary focus is on the reproducibility of two key tests: F-test and Levene's test. This exploration examines the relationship between RP values and the p-values and explores the relationship between RP in the rejection area and estimated power. For the F-test, the investigation extends to both two-sided testing and upper one-sided testing, and the examination of Levene's test focuses on the two-sided test. This is conducted through simulation studies, and the investigation takes into account scenarios involving both null and alternative hypotheses. The examination of RP of equality of variances tests is an important aspect of this thesis. The research focus is on understanding how a preliminary test of equality of variances influences the RP of location tests. Specifically, parametric location tests require the assumption of homogeneity, a critical aspect investigated through equality of variances tests.

Chapter 4 studies the reproducibility of the one-sample location tests with preliminary test (the two-stage procedure). This involves employing the Shapiro-Wilk test as the preliminary test for Normality to choose between the one-sample t-test or the one-sample Wilcoxon signed-rank test. The reproducibility of the two-stage procedure is assessed in various ways such as *Case A* which investigates full RP for all stages, *Case B* focuses on RP for the outcome of the location tests, and *Case C* examine RP of the location test conclusion, where for the bootstrap samples the same location test is applied as for the original sample. Furthermore, RP for the one-sample t-test and Wilcoxon test without performing the preliminary test is studied. Examine the impact of the preliminary test of Normality on RP of location tests by comparing RP for the location test with and without the preliminary test.

Chapter 5 addresses reproducibility for the two-sample location tests with preliminary tests (three-stage procedures). In the three-stage procedures, the Shapiro-Wilk test for Normality and the F-test for equality of two variances are used as preliminary tests to choose between location tests, the two-sample t-test, Welch's t-test, and the Wilcoxon-Mann-Whitney (WMW) test. The same cases in Chapter 4 are used to assess the RP. Additionally, RP for the two-sample t-test, Welch's t-test, and the WMW test without preliminary tests are studied. The effect of the preliminary tests on RP of the two-sample location tests is investigated by comparing the RP values for location tests with and without preliminary tests.

Chapter 6 presents reproducibility for the multiple-group location tests with preliminary tests. The focus on the RP for two and three-stage procedures for multiple-group location tests. The Shapiro-Wilk test is employed as a preliminary test for Normality in the two-stage procedure to choose between the one-way ANOVA and Kruskal-Wallis tests. The Shapiro-Wilk test and Levene's test are employed as preliminary tests for Normality and equality of variances in the three-stage procedure to choose between the

ANOVA test, Welch's ANOVA test and the Kruskal-Wallis test. The study is applied to 3 and 5 groups as examples for the multiple groups. Additionally, we explore RP for the location tests without preliminary tests to check the impact of the preliminary tests on RP for multiple-sample location tests by comparing RP for location tests with and without preliminary tests.

The final chapter 7 serves as a conclusion and the future works, summarizing key findings, highlighting contributions to the field, and suggesting directions for future research.

In this thesis, calculations were performed using the statistical software program R version 4.2.2. The random seed has been set to 1234 using 'set.seed(1234)' in the R codes to ensure that the same sequence of random numbers is generated each time the code is run, leading to identical samples.

Chapter 2

Reproducibility for Normality Tests

2.1 Introduction

The assumption of Normality plays a crucial role in many statistical analyses, as many procedures and tests rely on the assumption that the data follow a Normal distribution. Assessing Normality in statistical analysis is essential, as it enables researchers to establish if the data satisfies the assumption of Normality for specific statistical methods. Thus, the motivation of this chapter lies in investigating the reproducibility probability (RP) of Normality tests and understanding their role as preliminary tests in assessing the reproducibility of location tests in subsequent chapters.

There are plenty of types of Normality tests designed to assess whether a sample is drawn from a Normal population. The main tests for Normality include the Kolmogorov-Smirnov (KS) test [58] which evaluates the maximum difference between the empirical cumulative distribution function (CDF) of the sample and the theoretical CDF of the Normal distribution. This comparison enables the evaluation of how closely the sample data aligns with the distributional form of a Normal distribution. The Lilliefors corrected KS test [62] is another important Normality test. The Shapiro-Wilk [87] test and the Anderson-Darling (AD) [6] test are known for their robustness and sensitivity in detecting deviations from Normality, especially in smaller sample sizes. The Cramer-von Mises (CVM) test [31] evaluates the minimum distance between hypothetical and actual probability distribution. Finally, the D'Agostino test [32] evaluates the skewness of a sample distribution and helps in identifying substantial deviations from a symmetrical distribution.

This chapter focuses on three widely used Normality tests: the Shapiro-Wilk (SW) test, the Anderson-Darling (AD) test, and the Lilliefors (LF) test. These tests are particularly prominent in statistical software and are commonly employed for assessing the Normality assumption in various analyses. This chapter aims to assess RP for these Normality tests under the assumption of Normal distribution and distributions different from Normal distribution through simulation studies. Moreover, the chapter compares the RP values of these tests and investigates the relationship between their RP in the rejection area and their estimated powers. Additionally, this chapter examines the effect of significance level on the RP values of these Normality tests.

Section 2.2 provides a brief introduction to these three Normality tests. The simulation process is detailed in Section 2.3, where we simulate under both the null hypothesis and the alternative hypothesis for the Normality tests. The results of the simulation and the discussion of it are displayed in Subsection 2.3.1. Subsection 2.3.2 presents a comparison for reproducibility of these three Normality tests and discusses the relationship between their RP values in the rejection area and their estimated power. Subsection 2.3.3 compares RP values for different levels of significance. The chapter concludes by summarizing the most important results in Section 2.4.

2.2 Test for Normality

This section offers a brief introduction to three well-known Normality tests, namely the Shapiro-Wilk, Anderson-Darling, and Lilliefors tests. These tests are very sensitive to deviations from Normality and are widely recognized and employed for their robustness and effectiveness in assessing Normality across various sample sizes and distributions.

2.2.1 Shapiro-Wilk (SW) test

The Shapiro-Wilk (SW) test developed by Shapiro and Wilk [87], is an effective method for assessing deviations from Normality. The test statistic, denoted as W, is defined as the squared Pearson correlation coefficient between the order statistics of a sample and scores representing how the order statistics would appear if the population followed a Gaussian distribution [51]. Therefore, if the value of W is close to 1, the sample exhibits behaviour similar to a Normal distribution, if W is less than 1, the sample does not follow a Normal distribution [51].

The initial version of the SW test was originally designed for sample sizes ranging from $n = 3$ to 50. It included tabulated percentage points of the null distribution for p-values like 0.01, 0.02, 0.05, 0.1, 0.5, and others up to 0.99 [51]. Using tables was essential for calculating and interpreting the SW test according to Shapiro and Wilk [87]. They also suggested a normalizing transformation for W when $7 \leq n \leq 50$ to improve the test's power and accuracy by making the distribution of W closer to a standard Normal distribution [51]. However, even with this improvement, a reliance on tables remained necessary for $4 \leq n \leq 6$ because of the difficulty of deriving a simple mathematical formula for the W statistic and its distribution for these very small sample sizes [51]. Then in 1982, Royston [81] introduced an extension to the SW test that allowed larger sample sizes such as up to 2000; later in 1992 [82], he lifted this limitation even further up to 5000. Royston devised an approximate normalization method suitable for computer processing to compute the W value and its significance level for sample sizes ranging from $n = 3$ to 2000. Subsequently, an enhanced algorithm was introduced that encompassed sample sizes $3 \leq n \leq 5000$.

Assume the sample consists of n independent and identically distributed observations $X_1, X_2, \ldots, X_n$ from population X. The null hypothesis for the Shapiro-Wilk (SW) test is H_0: X is Normally distributed with unspecified μ and σ^2 ($N(\mu, \sigma^2)$), against the alternative hypothesis H_1: X is not Normally distributed. If $X_{(i)}$ for $i = 1, \ldots, n$ represents the n observations arranged in ascending sequence, the test statistic W is

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{2.2.1}$$

where n is the number of observations, and $\bar{X}$ is the sample mean. The vector of weights $\mathbf{a} = (a_1, \ldots, a_n)^T$ is determined by [82]:

$$\mathbf{a} = \frac{m^T V^{-1}}{(m^T V^{-1} V^{-1} m)^{1/2}} \tag{2.2.2}$$

where $m^T = (m_1, \ldots, m_n)$ is a vector composed of the expected values of the order statistics of n independent and identically distributed random variables sampled from the standard Normal distribution $(Y_{(n1)} \leq Y_{(n2)} \leq \ldots \leq Y_{(nn)})$, $E(Y_{(ni)}) = m_i$, and $V = v_{i,j}$ is the corresponding $n \times n$ covariance matrix of order statistics $\text{cov}(Y_{(ni)}, Y_{(nj)}) = v_{ij}$, for $i, j = 1, \ldots, n$ [75].

The null hypothesis for the SW test is rejected if $W \leq W_{(\alpha,n)}$ where α is the significance level. The critical value $W_{(\alpha,n)}$ can be obtained in Shapiro and Wilk [87]. The Shapiro-Wilk test is sensitive to deviations from Normality, especially in the centre part of the distribution. The Shpiro-Wilk test is strongest against short-tailed (platykurtic) and skew distributions and weakest against symmetric moderately long-tailed (leptokurtic) distribution [82]. Moreover, the Shapiro-Wilk test is sensitive to sample size; it tends to incorrectly reject Normality with large samples and fail to reject Normality with small samples [50].

2.2.2 Anderson-Darling (AD) test

The Anderson–Darling (AD) test for goodness-of-fit was first proposed in 1952 [5, 6]. The null hypothesis for the Anderson-Darling test applied to assess Normality is formulated as follows: H_0: data follow a Normal distribution. This is tested against the alternative hypothesis: H_1: data do not follow a Normal distribution. It involves ordering the test samples with size greater than 7 with unknown mean and variance, then calculating the statistic A^2:

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\{(2i-1)\log P_{(i)} + (2n+1-2i)\log(1-P_{(i)})\} \tag{2.2.3}$$

where $P_{(i)} = \Phi\left(\frac{X_{(i)}-\bar{X}}{\sigma}\right)$ is the cumulative distribution function (CDF) of the standard Normal distribution at $X_{(i)}$, and $X_{(i)}$ is the ordered sample of size n, $\bar{X}$ is the sample mean and σ is the standard deviation.

Stephens [94] suggests a modified statistic A^{2^*} to obtain critical values for all sample data:

$$A^{2^*} = A^2\left(1.0 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) \tag{2.2.4}$$

The p-values for the adjusted Anderson-Darling statistic can be calculated as follows: when $A^{2^*} \leq 0.2$, then p-value $= 1 - exp\left(-13.436 + 101.14A^{2^*} - 223.73(A^{2^*})^2\right)$. When

$0.2 < A^{2^*} \leq 0.34$, then p-value $= 1 - exp\left(-8.318 + 42.796A^{2^*} - 59.938(A^{2^*})^2\right)$. When $0.34 < A^{2^*} < 0.6$, then p-value $= exp\left(0.9177 - 4.279A^{2^*} - 1.38(A^{2^*})^2\right)$. Otherwise, then p-value $= exp\left(1.2937 - 5.709A^{2^*} + 0.0186(A^{2^*})^2\right)$.

The Anderson-Darling test is more sensitive to deviation in the tail than the central region [45]. This makes it more powerful for detecting deviations from Normality in the tails caused by skewness or heavy tails. This is because the weight function used in the Anderson-Darling test statistic gives more weight to the tails of the distribution than the central region.

2.2.3 Lilliefors (LF) test

The Lilliefors test examines whether the data are sampled from a Normal distribution with unknown mean and variance [62]. The initial step involves standardizing the data as $Y_i = \frac{X_i - \bar{X}}{S}$, where $\bar{X}$ is the sample mean and S is an unbiased estimate of the sample standard deviation [98]. Then, the test statistic D_n represents the maximum absolute difference between the empirical distribution function (EDF) of standardized data and that of a standard Normal distribution [98]:

$$D_n = \max_{i=1}^{n} \left| F_n(Y_i) - \Phi(Y_i) \right| \tag{2.2.5}$$

where $F_n(.)$ represents empirical distribution function and $\Phi(.)$ is CDF of the standard Normal distribution.

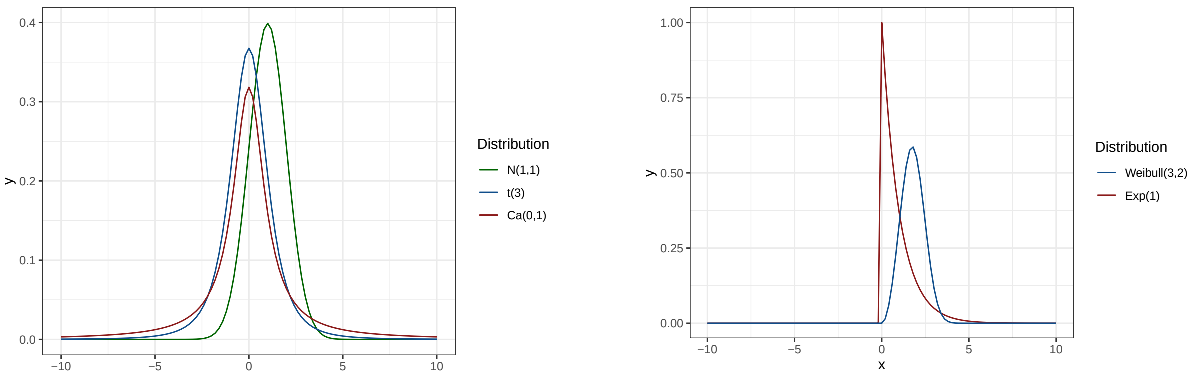
The null hypothesis of the sample following a Normal distribution with unknown mean and variance is rejected at the specified significance level α if the value D_n exceeds the $1 - \alpha$ quantile from the Lilliefors distribution tables [62].

The Lilliefors test has high sensitivity at the centre of the distribution [9]. The Lilliefors test has lower power compared to other Normality tests such as the Shapiro-Wilk or Anderson-Darling tests in identifying deviations from Normality [70].

2.3 Simulation studies for the reproducibility of the Normality tests

Simulation studies are conducted to investigate the reproducibility (RP) of the Shapiro-Wilk, Anderson-Darling and Lilliefors tests, where we use the NPI-B-RP Algorithm 1 that is presented in Section 1.4.5 of Chapter 1, to find the RP values for Normality tests. The inputs are an original sample with the sample size n, $h = 100$ and $N = 1000$.

Data are simulated under the null hypothesis of Normality tests H_0 using a Normal distribution with mean and variance equal to 1 denoted by $N(1, 1)$. Moreover, data are simulated under the alternative hypothesis of Normality H_1, where four different non-Normal distributions are considered, namely, the Weibull distribution with shape equal 3 and scale equal 2, denoted by $Weibull(3, 2)$, the Student's t-distribution with 3 degrees of freedom, denoted by $t(3)$, the Exponential distribution with rate 1, denoted by $Exp(1)$, and the Cauchy distribution with location parameter 0 and scale parameter 1, denoted by $Ca(0, 1)$. Non-Normal distributions were chosen to examine how different characteristics of the data distribution, such as skewness, long tails, heavy tails, and outliers, affect the reproducibility of the Normality tests. The Weibull distribution is chosen to examine how shape and scale characteristics impact Normality tests' reproducibility. The parameter of shape that equals 3 leads to a Weibull distribution that is slightly skewed to the right, and a scale parameter of 2 determines the width or spread of the distribution. The Student's t-distribution known for its relatively longer tail in comparison to the Normal distribution offers an understanding of how properties of data distribution, like tail behaviour and symmetry, affect the reproducibility of tests for Normality. The Exponential distribution, known for its long tail and constant hazard rate, provides a chance to investigate the reproducibility of Normality tests when the distribution of the original samples deviates substantially from the Normal distribution due to factors such as skewness, a long right tail, or outliers. Finally, the Cauchy distribution's heavy tails allow exploration of the reproducibility of the Normality tests in scenarios with extreme outliers and high variability. Figure 2.1 shows the shapes of the probability density functions (PDFs) of these chosen distributions.



(a) Symmetric distributions

(b) Asymmetric distributions

Figure 2.1: The probability density functions for selected distributions used in the simulation studies.

The number of runs per simulation is $K = 200$, for each run, one original sample of size n is generated from a chosen distribution, a Normality test is performed on this original sample and NPI-B-RP is calculated using Algorithm 1. Different n sample sizes, namely 5, 10, 20, and 50 are considered to evaluate the RP values across small, medium, and large sample sizes. Since the sample sizes of the Anderson-Darling test for Normality should be greater than 7 in the R program [46], we cannot apply the AD test for a sample size of 5. The tests are performed with a 5% level of significance and with a two-sided alternative hypothesis.

2.3.1 Results of the simulation studies

This subsection presents the results of the simulation studies for the reproducibility of the Shapiro-Wilk test, Anderson-Darling test and Lillifores test, under the null hypothesis for the Normality test when the distribution is $N(1,1)$, and under the alternative hypothesis when distributions are $Weibull(3,2)$, $t(3)$, $Exp(1)$, and $Ca(0,1)$.

Results under H_0

RP values for the Normality tests when data are simulated from $N(1,1)$ are presented in Figures 2.2 - 2.5 for sample sizes 5, 10, 20, and 50. In all Figures, the dotted line represents the threshold, with the rejection area to the left of the line and the non-rejection area to the right. The x-axis represents the p-values for the Normality tests, and the y-axis represents the minimum, mean and maximum NPI-B-RP values. It can be noticed that RP values for Normality tests tend to be high in the non-rejection area and low in the rejection area when dealing with small sample sizes. As the sample size grows, the opposite occurs, RP values tend to be larger for rejection of H_0 than for non-rejection. This is due to the power of Normality tests and the characteristics of the NPI-B samples. The NPI-B samples are diverse in distribution because the NPI-B does not make any distributional assumptions. Thus, when dealing with small sample sizes, these tests have lower power, making it more challenging to detect deviations from Normality. Therefore, the NPI-B samples can easily pass the Normality tests which lead to high RP in the non-rejection area and low RP in the rejection area. As the sample size increases, the tests become more powerful in identifying deviations from Normality. Therefore, the NPI-B samples have more difficulty passing the tests of Normality, Which results in lower RP values in the non-rejection area and higher RP values in the rejection area.

Furthermore, the relationship between RP values and p-values reveals a consistent pattern. Specifically, the RP value tends to be lower when the p-value is close to the threshold, gradually increasing as the p-value moves away from the threshold in both areas of rejection and non-rejection. This behaviour is rooted in the strength of evidence for or against the null hypothesis (H_0). When the p-value is close to the significance threshold, it indicates a marginal decision, where the evidence for or against H_0 is relatively weak. Consequently, the RP values of the tests are low because it is possible that the repeated experiment will not lead to the same result as the original test. As the p-value moves away from the threshold, the evidence for or against H_0 strengthens, resulting in higher RP values. This relationship has been observed in various studies such as [2, 13, 91].

It can be seen that as p-values for the Normality tests are close to one, the values of RP of the Normality tests remain low and do not approach one. This is because the simulations study is performed for the two-sided Normality tests. The two-sided tests are designed to detect deviations from the null hypothesis in both directions. This means they consider the possibility of deviations in either tail of the distribution. As a result, even when the p-value is close to one, indicating weak evidence against the null hypothesis, there might still be doubt regarding the actual underlying distribution. Consequently,

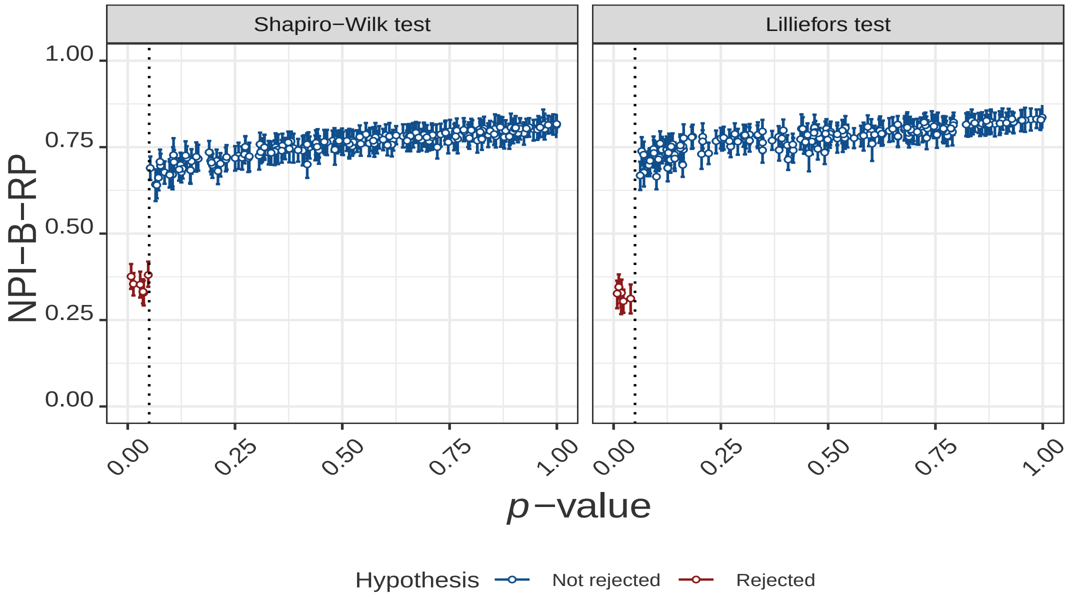


Figure 2.2: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from $N(1, 1)$, with $n = 5$, $\alpha = 0.05$

the RP values do not approach one.

As the sample size increases, the variability in RP values of all tests increases. The variability of the RP values attributed to the characteristics of the NPI-B samples and the behaviour of different Normality tests with data. As the sample size increases, the greater complexity and diversity of the NPI-B samples lead to a wider range of potential test outcomes, resulting in increased variability in the RP values. The Anderson-Darling test shows lower variability in RP values when compared to the Shapiro-Wilk test and the Lilliefors test. This difference may stem from the greater sensitivity of the Anderson-Darling test to deviations in the tails than the median, making it more powerful for detecting deviations from Normality in the tails. This led to more consistent results among different NPI-B samples, resulting in decreased variability in RP. In contrast, the Shapiro-Wilk and Lilliefors tests are often more sensitive to departures in the central part of the distribution. Also, the Anderson-Darling test tends to be less sensitive to sample size variations compared to other tests, which may lead to more stable results and lower variability in RP.

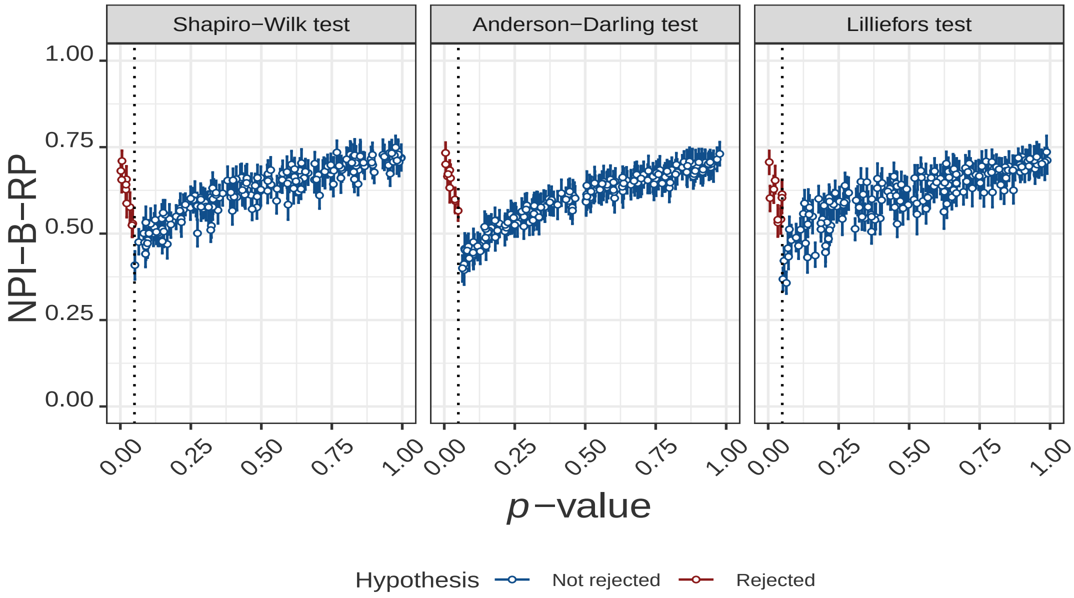


Figure 2.3: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $N(1, 1)$, with $n = 10$, $\alpha = 0.05$

Results under H_1

This subsection presents the results of the simulation studies examining RP values for the Normality tests when the data are sampled from distributions that differ from a Normal distribution, namely $Weibull(3, 2)$, $t(3)$, $Exp(1)$ and $Ca(0, 1)$ distributions. These distributions are chosen to provide insight into how deviations from Normality, such as heavy tails, skewness, and other non-Normal characteristics, impact the RP of Normality tests.

Figures 2.6 - 2.9 show RP values for the Normality test when data are drawn from the Weibull distribution. Generally, RP values exhibit patterns close to observed when sampling from $N(1, 1)$, presented in Subsection 2.3.1, in terms of the relationship between RP values and p-values for the Normality tests, the effect of sample size in RP values and variability in the RP values. Although, $Weibull(3, 2)$ is non-Normal distribution, the number of original samples that reject and non-reject H_0 is close to those when data sampled from $N(1, 1)$, as shown in Tables 2.1 and 2.2. This may be because when the shape parameter is equal to 3, the Weibull distribution approximates the Normal distribution [71].

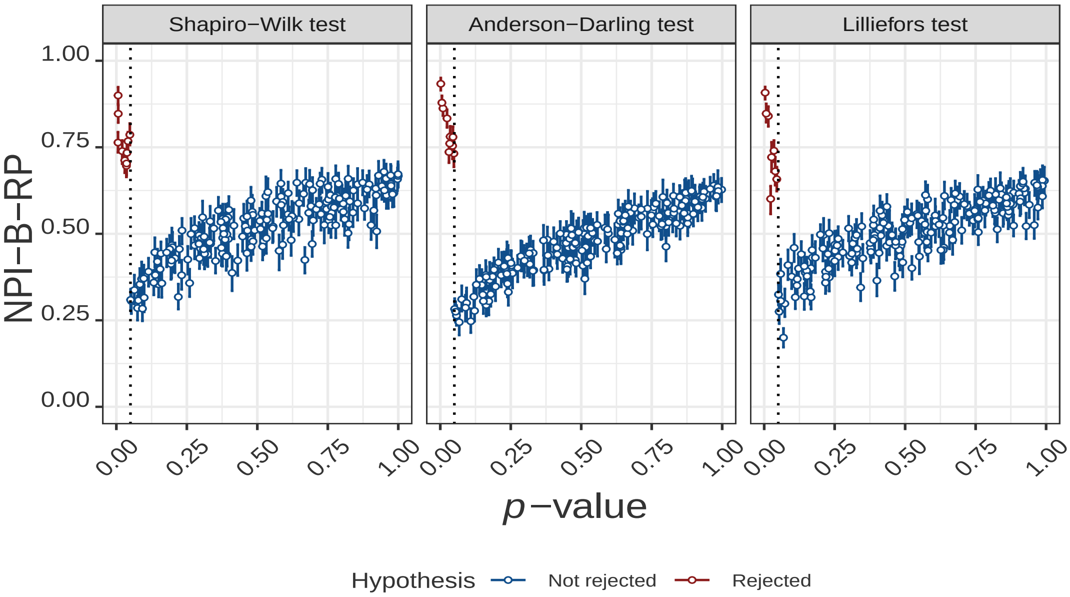


Figure 2.4: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $N(1,1)$, with $n = 20$ $\alpha = 0.05$

Figures 2.10 - 2.13 show the results for RP of the Normality tests when data are sampled from $t(3)$. It can be seen that the RP values exhibit similar patterns observed when sampling from $N(1,1)$, presented in Subsection 2.3.1, in terms of the relationship between RP values and p-values for the Normality tests, the effect of sample size in RP values and variability in the RP values. However, the number of original samples in the rejection region that increases with increasing sample size is higher than that when the data are sampled from Normal and Weibull distributions, as given in Table 2.3.

Figures 2.14 - 2.17, show the results of RP for the Normality tests when data are sampled from $Exp(1)$ which is a long right tail distribution. RP values show similar patterns observed when sampling from $t(3)$. However, the number of original samples that are located in the rejection area is higher than that observed in the $t(3)$ distribution case, as shown in Table 2.4. This is because the long right-tail of the Exp distribution results in an increase in extreme values, outliers, and skewness in the data. As a result, this impacts the performance of tests for Normality. Moreover, it is observed that RP values in the rejection area tend to be high, and as the sample size increases, RP values approach close to one. Whereas in the non-rejection area, as the sample size increases

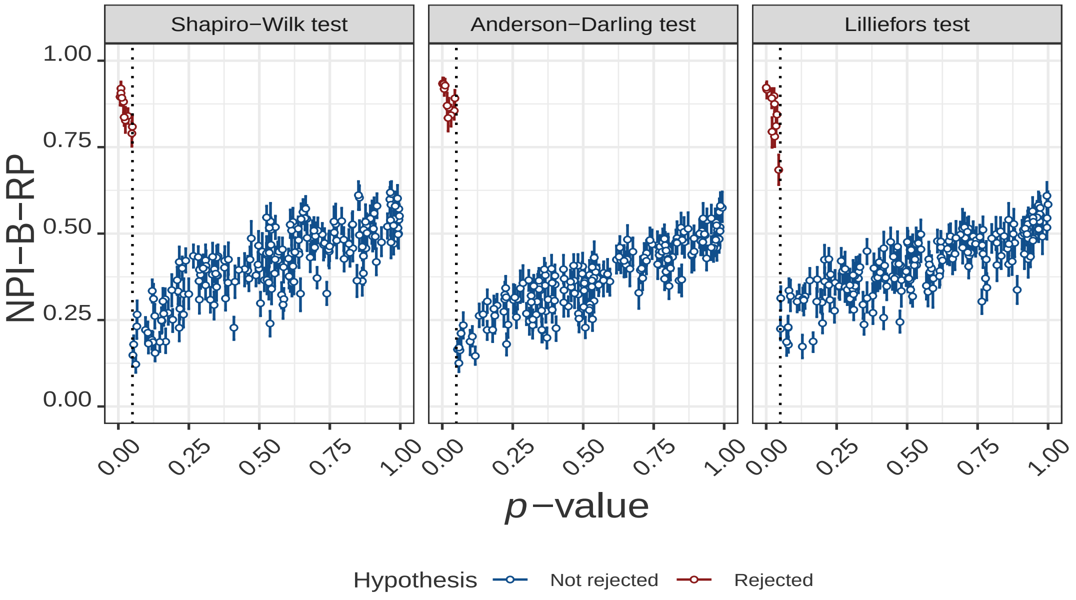


Figure 2.5: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $N(1, 1)$, with $n = 50$, $\alpha = 0.05$

the number of original samples becomes small and has p-values close to the threshold with low RP. This is simply due to the increase in the test power for larger sample sizes. Therefore, RP decreases relatively for most of the non-rejection cases with larger sample sizes compared to the non-rejection cases with smaller sample sizes.

Figures 2.18 - 2.21 show RP values when original samples are drawn from $Ca(0, 1)$ which is characterized by a heavy tail, the RP values for the Normality tests appear similar patterns observed when sampling from $Exp(1)$. However, the number of original samples in the rejection area and RP values in the rejection area for all the Normality tests are higher than observed in $Exp(1)$. This can be traced back to the heavy-tailed nature of the Cauchy distribution which leads to a higher probability of extreme values occurring in the original samples and then the NPI-B samples. Thus, these extreme values contribute to more original samples located in the rejection area and then high RP values. Conversely, the Exponential distribution demonstrates different tail behaviour compared to the Cauchy distribution. Although it also deviates from Normality, its tail behavior may not be as pronounced as that of the Cauchy distribution leading to a comparatively smaller number of original samples located in the rejection area and the low RP values compared to $Ca(0, 1)$.

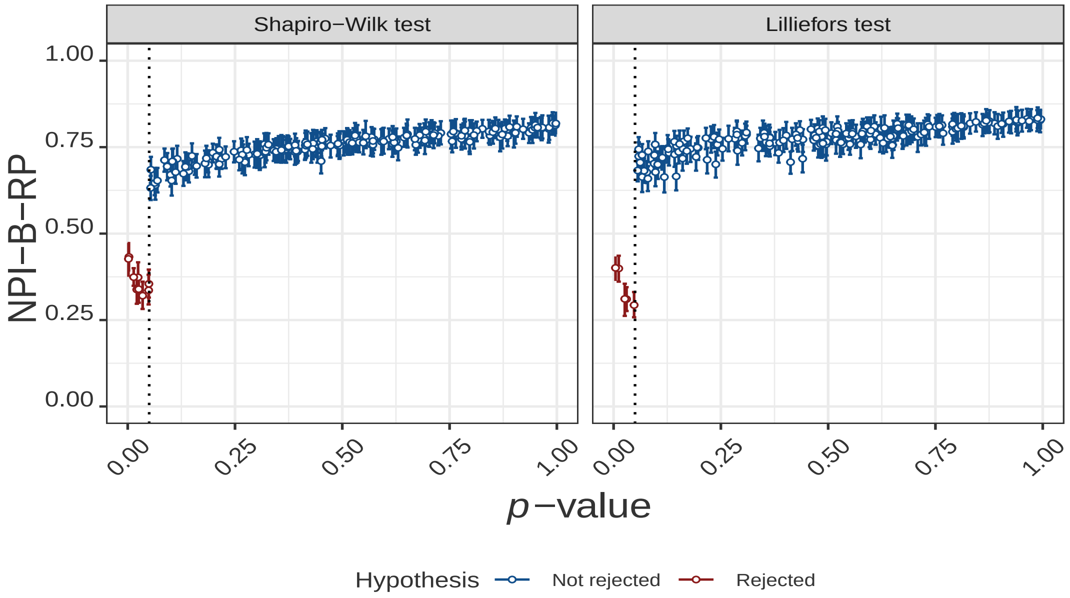


Figure 2.6: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors tests for data sampled from $Weibull(3,2)$, with $n = 5$, $\alpha = 0.05$

2.3.2 Comparison between the Normality tests

A comparison is conducted between the Shapiro-Wilk test, Anderson-Darling test, and Lilliefors test based on the overall mean of RP values (which are computed by averaging the mean RP values across different original samples) and the number of original samples in the rejection area. Furthermore, the relationship between the overall mean of RP values in the rejection area with the estimated power for the Normality tests is examined. Generally, there is no substantial difference between the Normality tests in their reproducibility and their estimated power.

Table 2.1 shows the number of original samples that are located in the non-rejection and rejection areas and the overall mean of their RP values when data is from $N(1,1)$. The observed proportions of cases where H_0 is falsely rejected are in line with this predetermined $\alpha = 0.05$. Additionally, it can be seen that the overall mean of RP values for the Shapiro-Wilk test and Lilliefors test are approximately similar for sample sizes 10, 20, and 50, and their RP values are higher than the RP values for the Anderson-Darling test in the non-rejection area. While in the rejection area, RP for the Anderson-Darling test is slightly higher than RP for the Shapiro-Wilk and Lilliefors tests. For the sample size

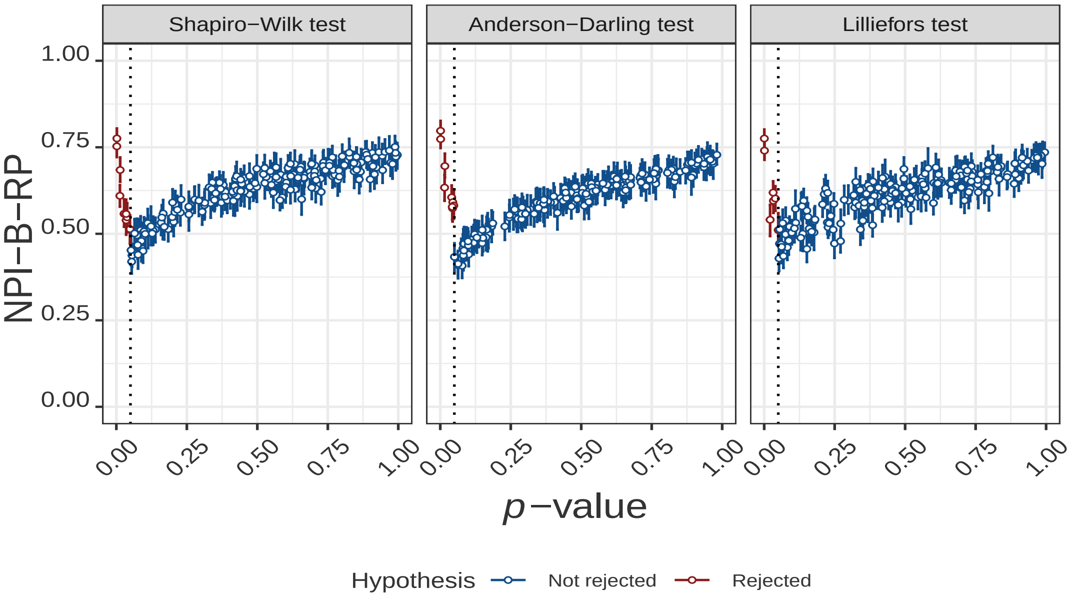


Figure 2.7: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Weibull(3, 2)$, with $n = 10$, $\alpha = 0.05$

of 5, The Lilliefors test has a higher overall mean of RP values in the non-rejection area and lower in the rejection area than the RP of the Shapiro-Wilk test.

In the case of sampling under H_1 which is considered in Subsection 2.3.1. When data are drawn from $Weibull(3, 2)$, Table 2.2 reveals that the Anderson-Darling test shows a slightly higher number of original samples that reject (H_0). For a sample size of 5, the number of original samples that reject (H_0) for the Shapiro-Wilk test is slightly higher than the Lilliefors test. Moreover, the Anderson-Darling test has the highest overall mean of RP values in the rejection area, while in the non-rejection area, the Shapiro-Wilk test has the highest RP for sample sizes 10, 20, and 50. For size 5, the Lilliefors test has a higher overall mean of RP values in the non-rejection area and has a lower RP in the rejection area compared to the Shapiro-Wilk test.

In scenarios where the alternative distribution is $t(3)$ a symmetric distribution with longer tails than the Normal distribution, from the results presented in Table 2.3, it is observed that the Shapiro-Wilk test shows a relatively highest number of original samples that reject (H_0). In contrast, the Lilliefors test shows the least number of rejections. Notably, the Shapiro-Wilk test shows a higher mean of RP values in the non-rejection

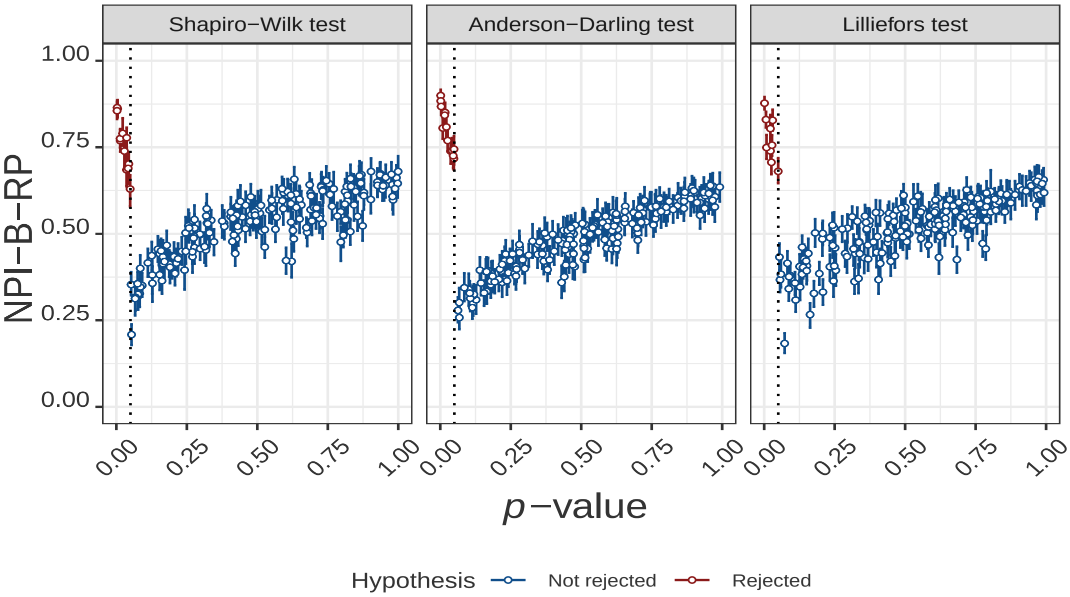


Figure 2.8: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Weibull(3, 2)$, with $n = 20$, $\alpha = 0.05$

case for sample sizes 10, 20, and 50. Conversely, the overall mean of RP values of the Anderson-Darling test is the highest in the rejection case. For a sample size of 5, the Lilliefors test has a higher RP in the non-rejection area and lower in the rejection area compared to the RP of the Shapiro-Wilk test.

Table 2.4 presents the number of original samples and the overall mean of RP values for the Normality tests in both areas when data is sampled from the $Exp(1)$ distribution. The number of original samples that reject (H_0) for the Shapiro-Wilk test is the highest compared to other tests. However, it is worth noting that the number of original samples that reject (H_0) of the Shapiro-Wilk and the Anderson-Darling tests is similar for the sample size of 50. Conversely, the Lilliefors test shows the least number of rejections. Notably, the Shapiro-Wilk test has a little bit higher RP in the non-rejection area for sample sizes 10 and 20, while the Anderson-Darling test has the highest RP in the rejection area for sample sizes 10 and 20, followed by the Lilliefors test. For a sample size of 50, the Shapiro-Wilk test and the Anderson-Darling test have the same mean of RP values in the rejection area. For a sample size of 5, the Lilliefors test has a higher mean of RP values in the non-rejection case and lower in the rejection area compared to the Shapiro-Wilk test.

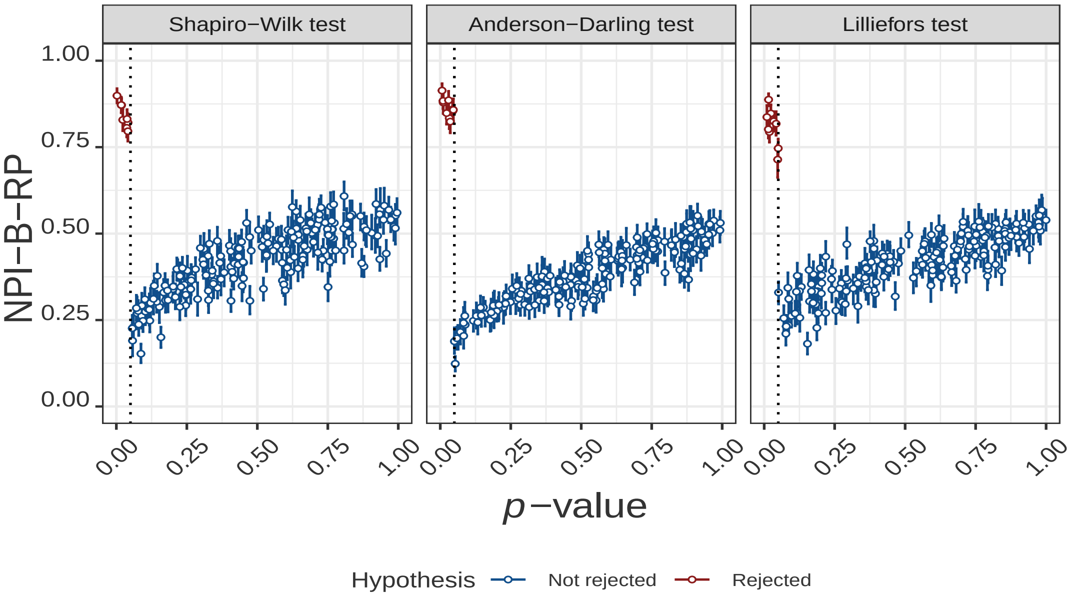


Figure 2.9: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Weibull(3, 2)$, with $n = 50$, $\alpha = 0.05$

Table 2.5 shows RP and the number of original samples for the Normality tests when data are sampled from a heavy-tailed symmetric distribution $Ca(0, 1)$. The Anderson-Darling test exhibits a slightly higher number of original samples that reject (H_0) than the Shapiro-Wilk test for sample sizes 10 and 20, while this number is similar for both tests for sample size $n = 50$. Conversely, the Lilliefors test consistently demonstrates the lowest number of rejections for sample sizes 10, 20, and 50. In terms of RP values, the Anderson-Darling test has a slightly higher RP in the rejection case for sample sizes $n = 10$ and $n = 20$, followed by the Lilliefors test. For the sample size of 50, all tests have approximately the same RP in the rejection area. whereas in the non-rejection area, the Shapiro-Wilk test has slightly higher RP than other tests. For the sample size of 5, the Lilliefors test has a higher RP in the non-rejection case and lower in the rejection area compared to the RP of the Shapiro-Wilk test.

The estimated power of Normality tests is evaluated through Monte Carlo simulations of 100, 000 datasets for non-Normal distributions $Weibull(3, 2)$, $t(3)$, $Exp(1)$, and $Ca(0, 1)$, for different sample sizes of 10, 20, and 50. The Monte Carlo simulations are calculated as the proportion of times a test produces p-value below the significance level $(\alpha = 0.05)$ when the H_1 of non-Normality is true to the total number of the simulation.

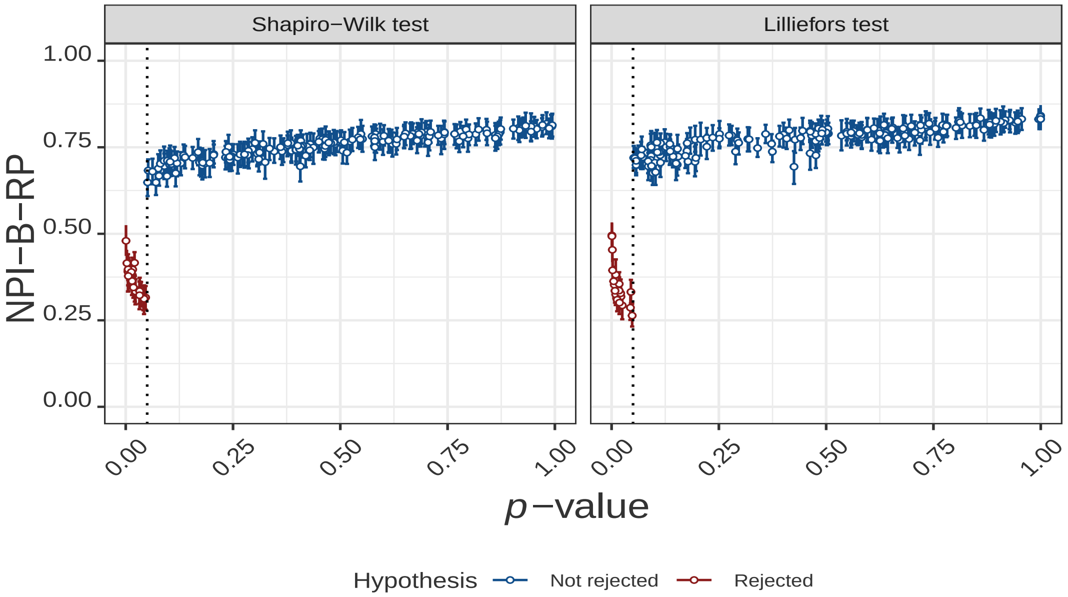


Figure 2.10: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from $t(3)$, with $n = 5$, $\alpha = 0.05$

The results of simulations for the estimated power of Normality tests are shown in Figure 2.22. The results show that the estimated power for the Normality test increases as the sample size increases. Furthermore, from the results the Shapiro-Wilk and Anderson-Darling tests approximately show similar power. Whereas, the Lilliefors test demonstrates the lowest power, particularly for $t(3)$ and $Exp(1)$ distributions.

When comparing the estimated power of Normality tests with the overall mean of RP values in the rejection area, we generally observe that as the estimated power of the Normality tests increases, the RP in the rejection area also increases. When compared between tests, although the Shapiro-Wilk test has slightly higher power than Anderson-Darling for $t(3)$ and $Exp(1)$ distributions, Anderson-Darling has the highest RP in the rejection area, except for the sample size $n = 50$ and distribution $Exp(1)$ both Shapiro-Wilk and Anderson-Darling have the same power and RP. The Anderson-Darling exhibited higher power and RP for the $Ca(0, 1)$ distribution, especially for sample sizes of 10 and 20. Interestingly, although the Lilliefors test has the lowest power, it does not consistently yield the lowest RP in the rejection area across all scenarios.

All tests showed greater estimated power when used on data from the Cauchy $(0, 1)$ dis-

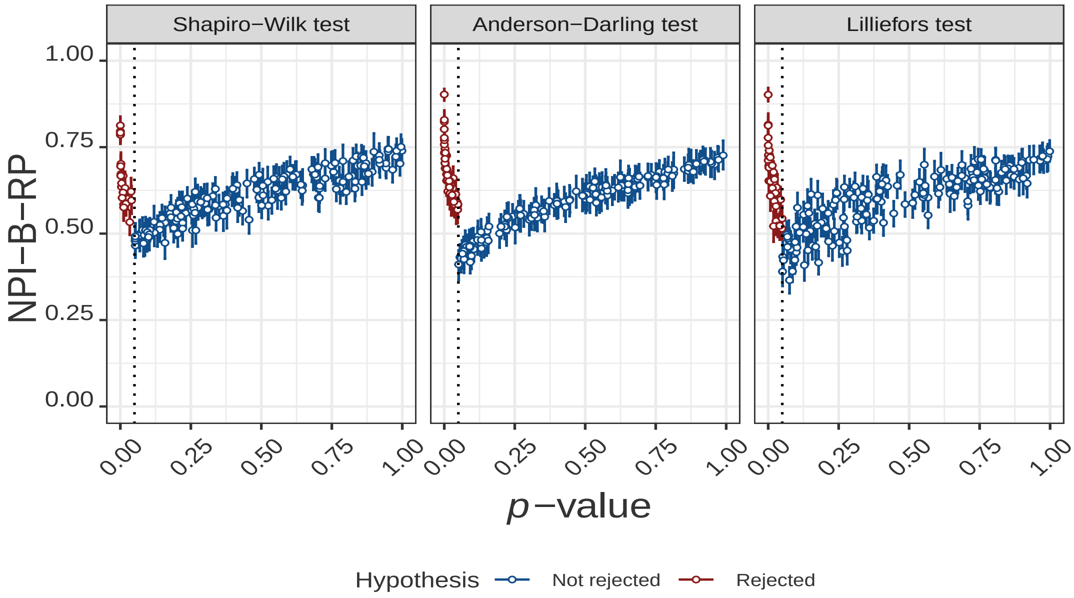


Figure 2.11: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $t(3)$, with $n = 10$, $\alpha = 0.05$

tribution and have the highest overall mean of RP values in the rejection area. Conversely, these tests have the smallest estimated power when used on data with $Weibull(3, 2)$ distribution and have the lowest mean of RP values in the rejection area. This is because the Normality tests have high power to detect departures from Normality in data sampled from a heavy-tailed distribution. When dealing with data that follow $Weibull(3, 2)$ distribution with shape parameter 3 the distribution approaches Normal distribution, this makes it difficult for Normality tests to detect deviations from Normality.

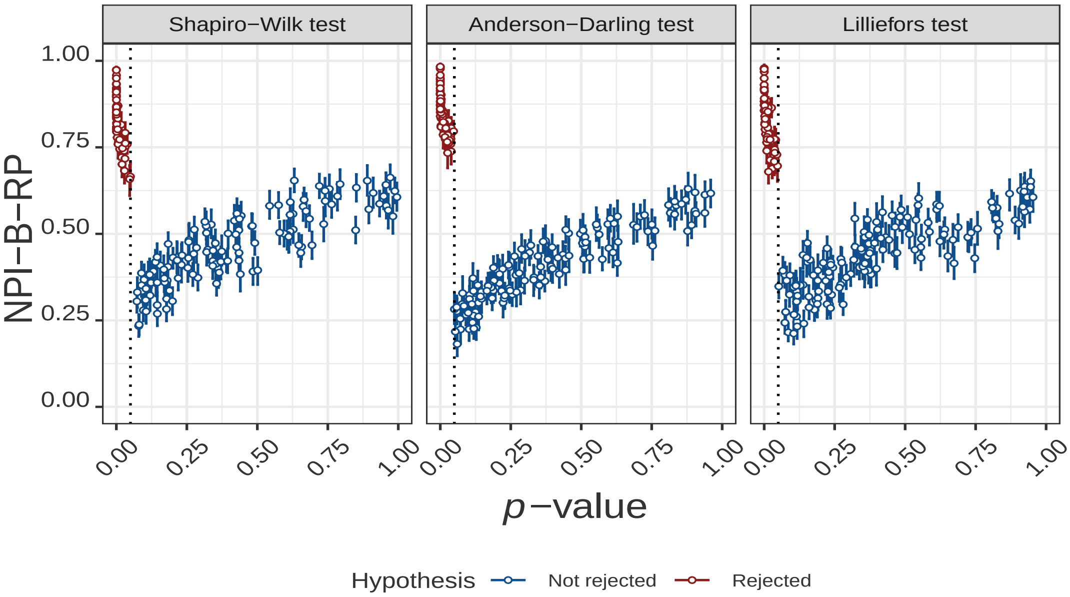


Figure 2.12: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $t(3)$, with $n = 20$, $\alpha = 0.05$

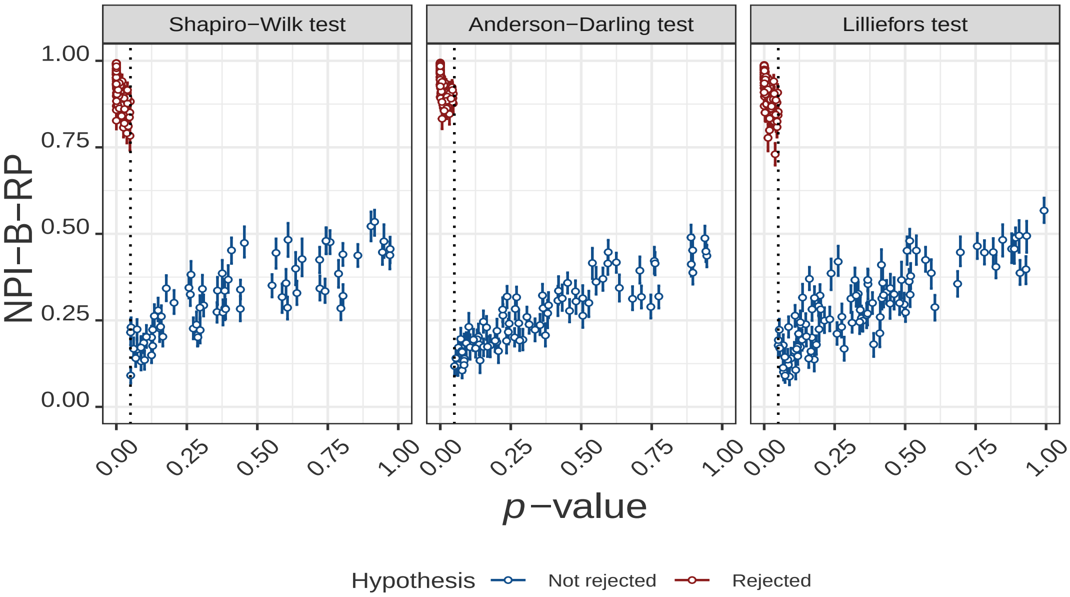


Figure 2.13: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $t(3)$, with $n = 50$, $\alpha = 0.05$

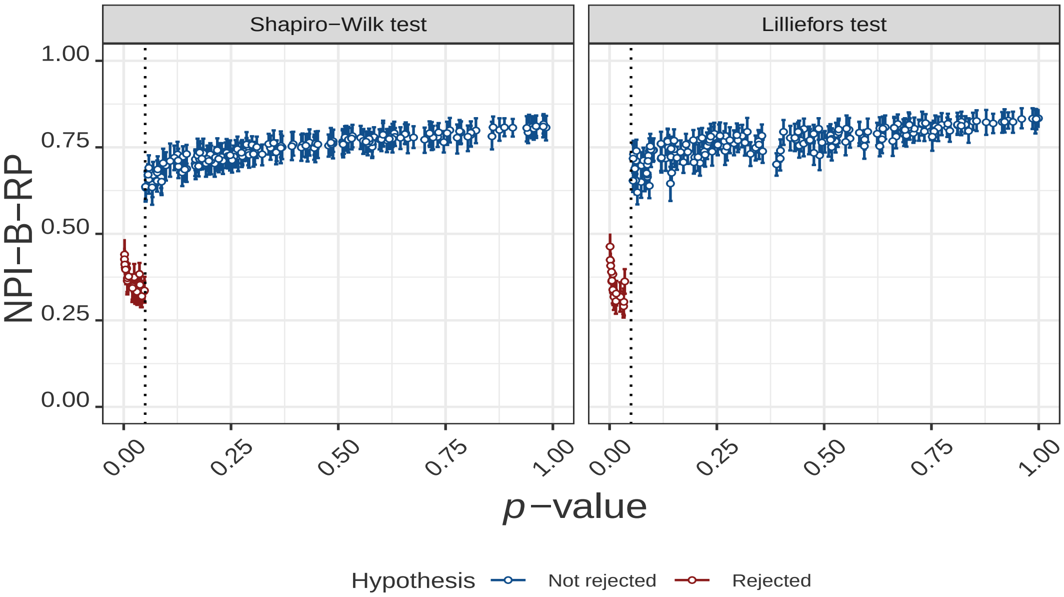


Figure 2.14: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from $Exp(1)$, with $n = 5$, $\alpha = 0.05$

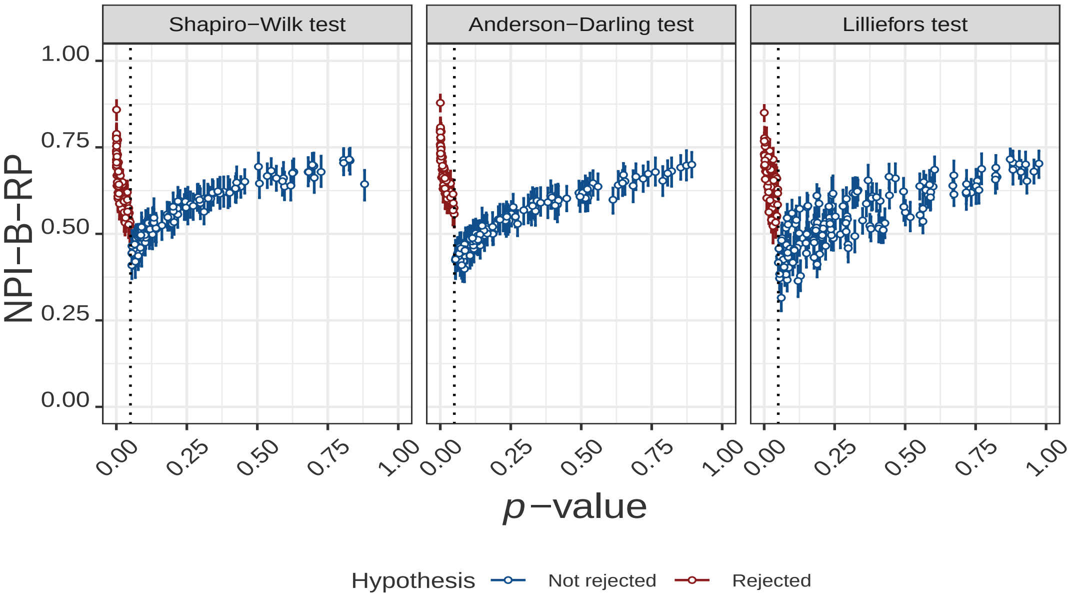


Figure 2.15: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Exp(1)$, with $n = 10$, $\alpha = 0.05$

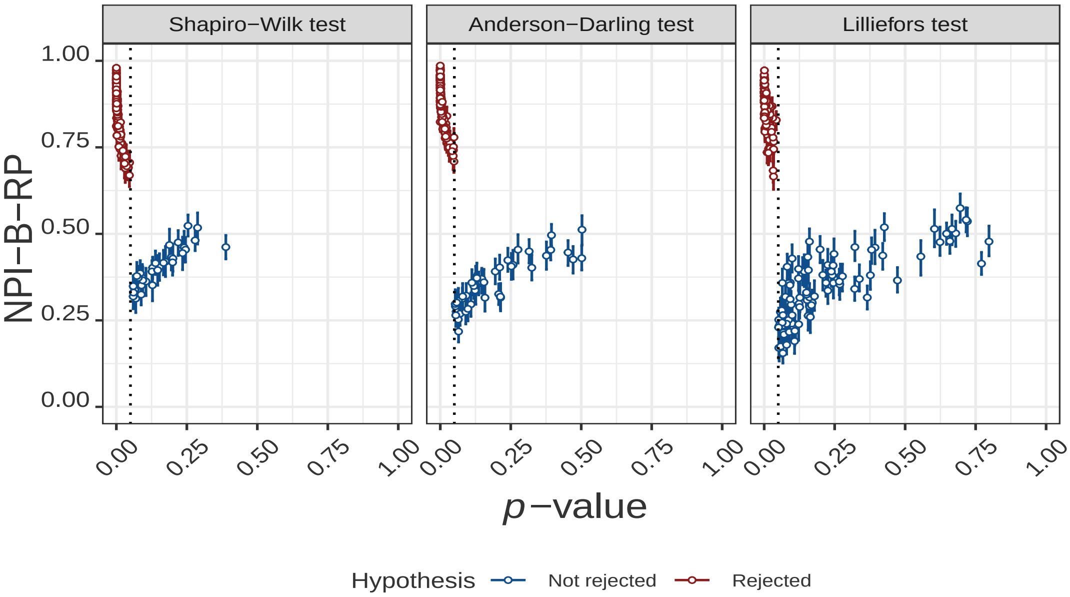


Figure 2.16: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Exp(1)$, with $n = 20$, $\alpha = 0.05$

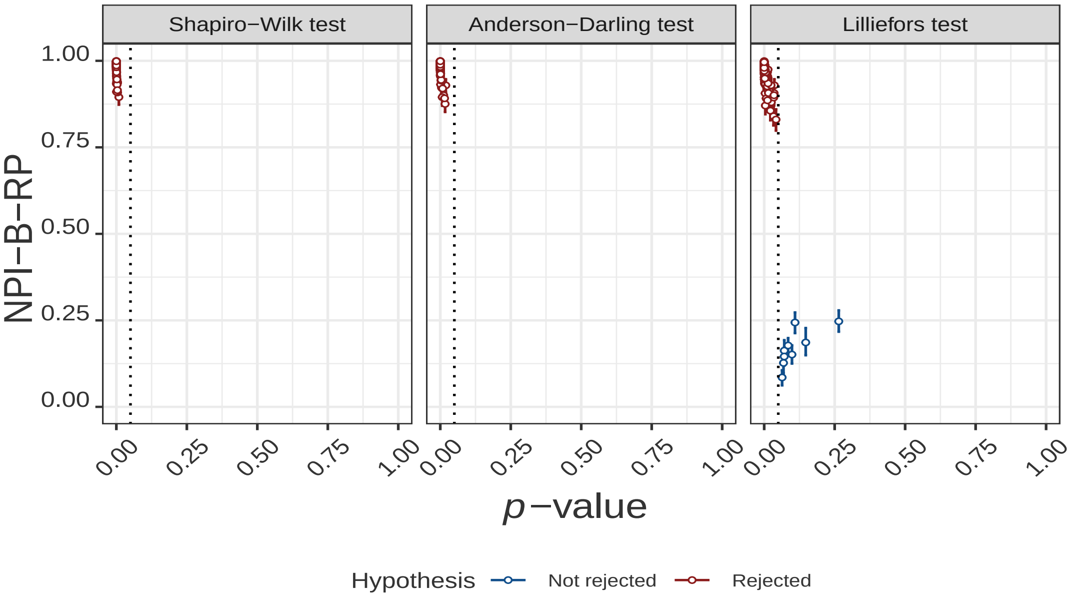


Figure 2.17: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Exp(1)$, with $n = 50$, $\alpha = 0.05$

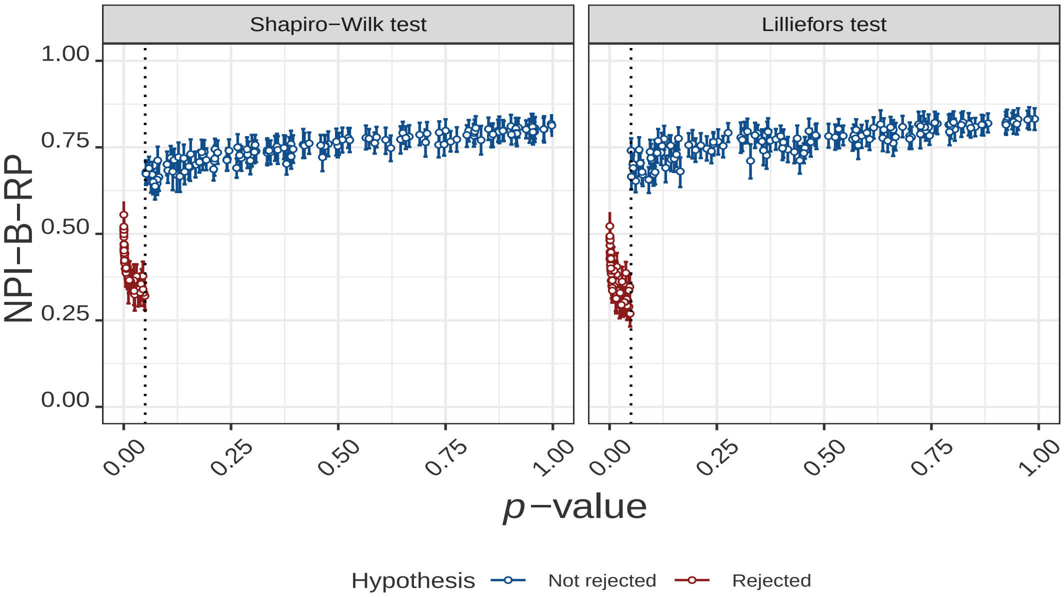


Figure 2.18: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from $Ca(0, 1)$, with $n = 5$, $\alpha = 0.05$

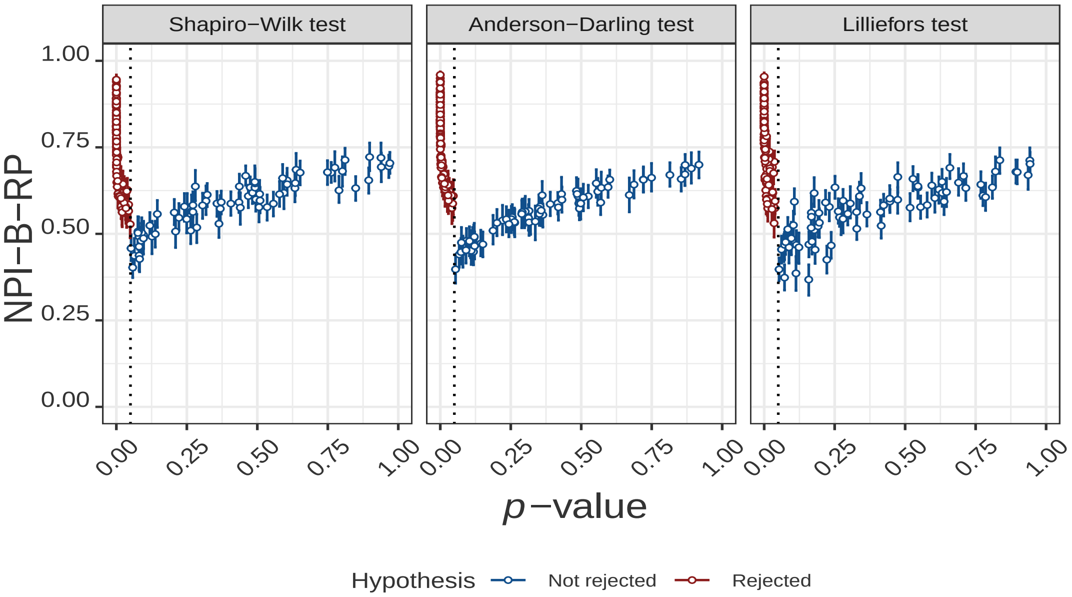


Figure 2.19: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Ca(0, 1)$, with $n = 10$, $\alpha = 0.05$

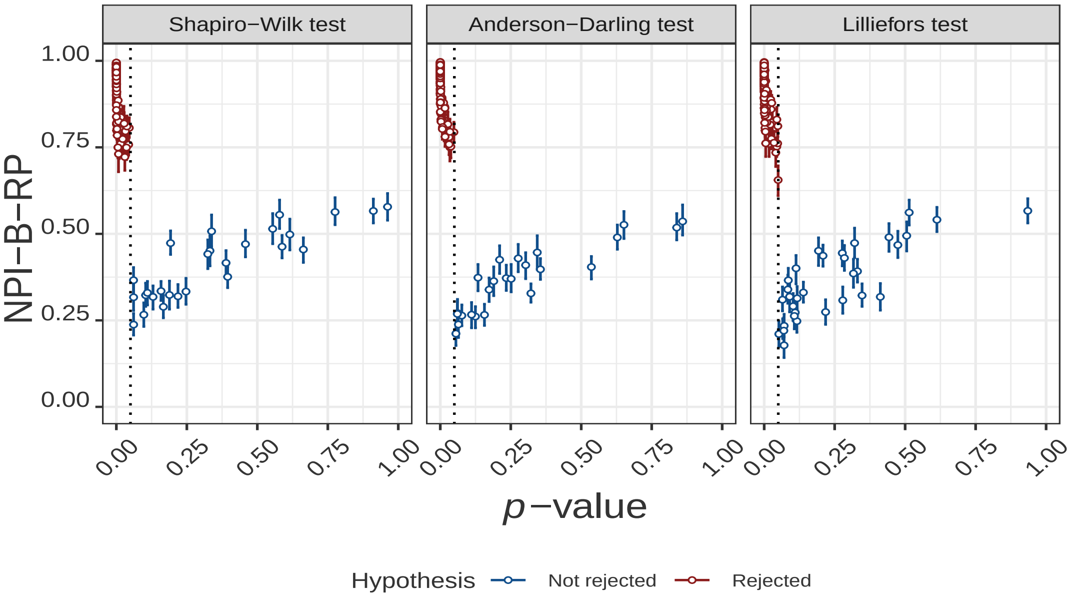


Figure 2.20: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Ca(0,1)$, with $n = 20$, $\alpha = 0.05$

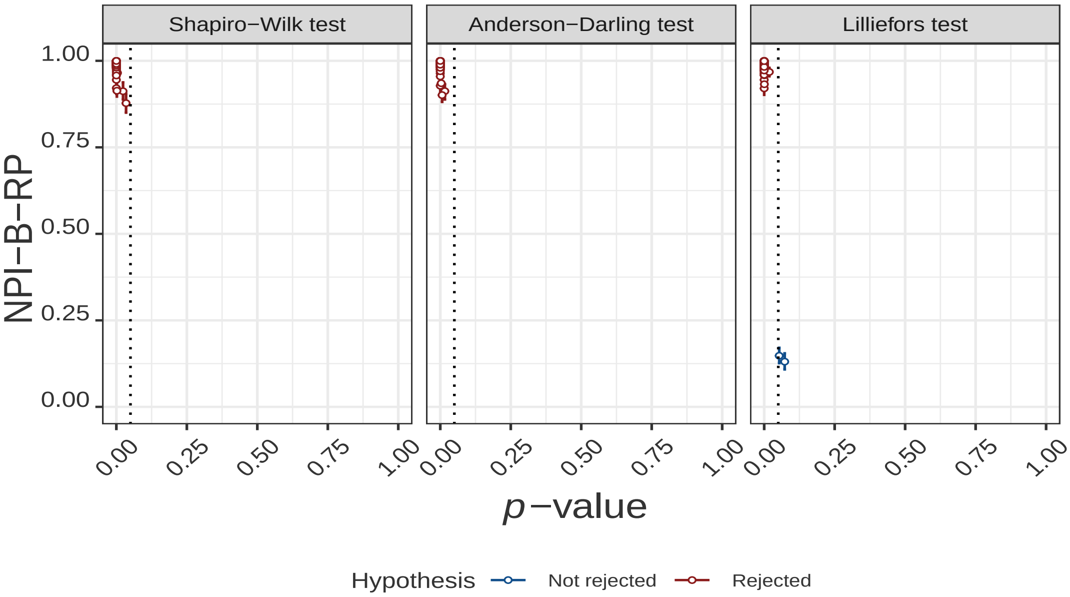


Figure 2.21: The relationship between NPI-B-RP and p-value for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Ca(0,1)$, with $n = 50$, $\alpha = 0.05$

Tests	Shapiro-Wilk				Anderson-Darling				Lilliefors			
	Rejection		Non-rejection		Rejection		Non-rejection		Rejection		Non-rejection	
	R	RP	N	RP	R	RP	N	RP	R	RP	N	RP
$n = 5$	6	0.354	194	0.759	-	-	-	-	6	0.320	194	0.781
$n = 10$	10	0.619	190	0.630	10	0.650	190	0.602	9	0.602	191	0.620
$n = 20$	10	0.764	190	0.524	10	0.805	190	0.477	8	0.749	192	0.509
$n = 50$	10	0.859	190	0.418	11	0.891	189	0.375	10	0.841	190	0.413

Table 2.1: Rejection (R) and non-rejection (N) counts for original samples from $N(1,1)$, along with the mean of RP values.

Tests	Shapiro-Wilk				Anderson-Darling				Lilliefors			
	R	RP	N	RP	R	RP	N	RP	R	RP	N	RP
$n = 5$	**9**	**0.366**	191	0.757					5	0.342	195	0.776
$n = 10$	**9**	0.615	191	0.631	**9**	**0.647**	191	0.604	7	0.626	193	0.622
$n = 20$	12	0.761	188	0.535	**13**	**0.799**	187	0.4908	9	0.774	191	0.518
$n = 50$	8	0.840	192	0.427	**10**	**0.864**	190	0.382	9	0.807	191	0.417

Table 2.2: Rejection (R) and non-rejection (N) counts for original samples from $Weibull(3,2)$, along with the mean of RP values.

Tests	Shapiro-Wilk				Anderson-Darling				Lilliefors			
	Rejection		Non-rejection		Rejection		Non-rejection		Rejection		Non-rejection	
	R	RP	N	RP	R	RP	N	RP	R	RP	N	RP
$n = 5$	**20**	**0.364**	180	0.755	-	-	-	-	19	0.353	181	0.777
$n = 10$	**52**	0.659	148	0.609	47	**0.670**	153	0.588	36	0.658	164	0.596
$n = 20$	**68**	0.824	132	0.458	67	**0.849**	133	0.409	57	0.822	143	0.438
$n = 50$	**126**	0.925	74	0.309	116	**0.939**	84	0.264	98	0.917	102	0.286

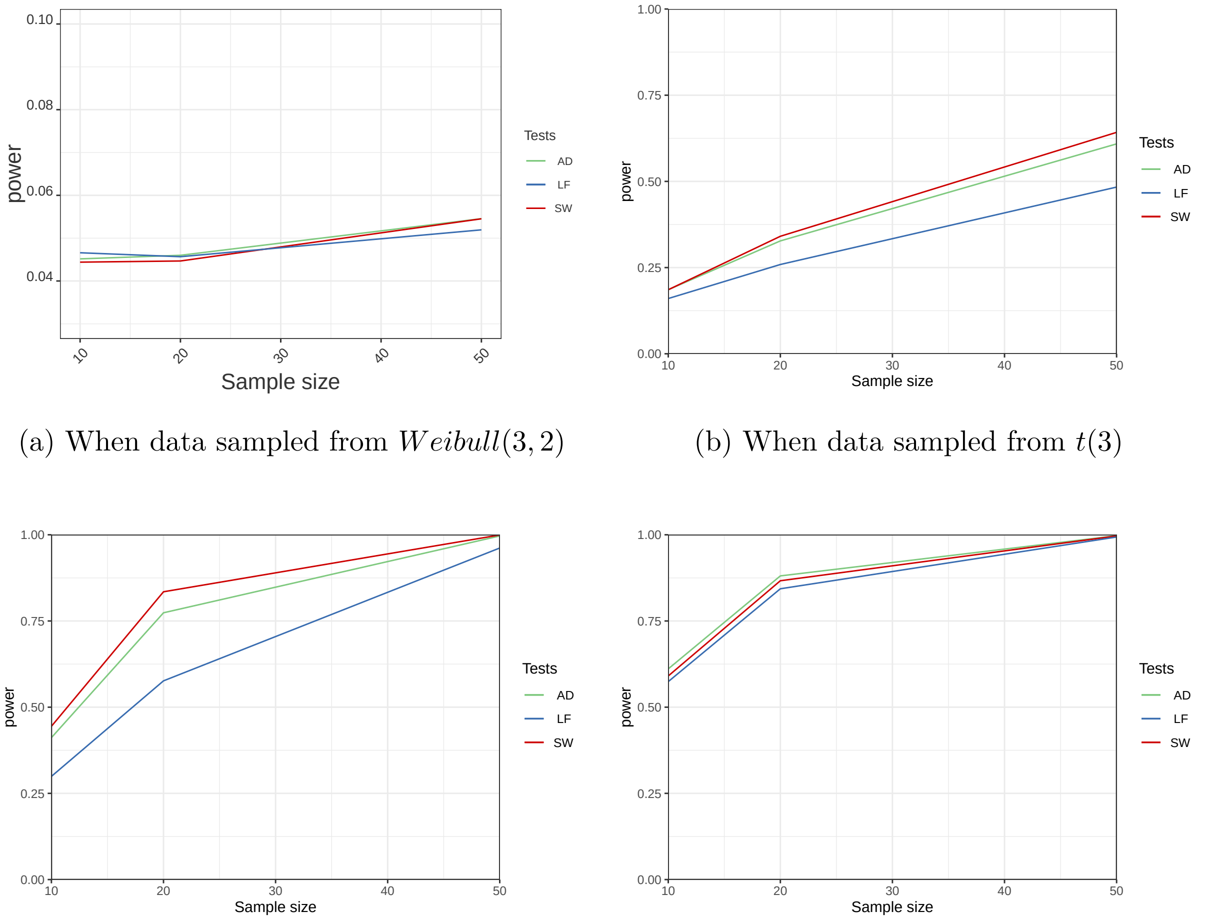
Table 2.3: Rejection (R) and non-rejection (N) counts for original samples from $t(3)$, along with the mean of RP values.

Tests	Shapiro-Wilk				Anderson-Darling				Lilliefors			
	Rejection		Non-rejection		Rejection		Non-rejection		Rejection		Non-rejection	
	R	RP	N	RP	R	RP	N	RP	R	RP	N	RP
$n = 5$	**29**	**0.363**	171	0.747	-	-	-	-	17	0.352	183	0.764
$n = 10$	**85**	0.652	115	0.559	78	**0.682**	122	0.533	51	0.655	149	0.540
$n = 20$	**166**	0.854	34	0.402	155	**0.875**	45	0.355	105	0.856	95	0.341
$n = 50$	**200**	**0.984**	0	N/A	**200**	**0.984**	0	N/A	191	0.966	9	0.169

Table 2.4: Rejection (R) and non-rejection (N) counts for original samples from $Exp(1)$, along with the mean of RP values.

Tests	Shapiro-Wilk				Anderson-Darling				Lilliefors			
	Rejection		Non-rejection		Rejection		Non-rejection		Rejection		Non-rejection	
	R	RP	N	RP	R	RP	N	RP	R	RP	N	RP
$n = 5$	66	**0.411**	134	0.748	-	-	-	-	**67**	0.380	133	0.769
$n = 10$	124	0.753	76	0.586	**128**	**0.777**	72	0.561	122	0.767	78	0.567
$n = 20$	173	0.922	27	0.410	**176**	**0.935**	24	0.3658	167	0.927	33	0.359
$n = 50$	**200**	**0.993**	0	N/A	**200**	**0.993**	0	N/A	198	0.991	2	0.139

Table 2.5: Rejection (R) and non-rejection (N) counts for original samples from $Ca(0, 1)$, along with the mean of RP values.



(a) When data sampled from $Weibull(3, 2)$

(b) When data sampled from $t(3)$

(c) When data sampled from $Exp(1)$

(d) When data sampled from $Ca(0, 1)$

Figure 2.22: The estimated power for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $Weibull(3, 2)$, $t(3)$, $Exp(1)$, and $Ca(0, 1)$, with sample sizes 10, 20, and 50.

2.3.3 Reproducibility of the Normality tests for different levels of significance

The same simulation studies presented in Section 2.3 were conducted under H_0 for the Normality tests, with different levels of significance, namely 0.1 and 0.01. As a result, we observe that from Figures 2.23 - 2.28, the RP values follow the same general pattern which is that RP values tend to be low when p-values are close to a level of significance and when p-values are far away from a level of significance RP values are high.

In the non-rejection area, RP values are generally high at lower significance levels and decrease as the level of significance increases. Conversely, in the rejection area, RP values tend to be notably low at small significance levels and increase with higher significance levels. This is because if α decreases, it becomes more stringent and one tends to reject H_0 in fewer cases, leading to high RP in the non-rejection area and low RP in the rejection area. The opposite happens at a high level of significance.

Figure 2.29 illustrates the overall mean of RP values for the Normality tests under H_0, with significance levels α set at 0.01, 0.05, and 0.1. For sample sizes of 5, the LF test exhibits slightly higher RP values than the SW test in the non-rejection area when α is 0.01 and 0.05. However, at $\alpha = 0.1$, both tests show approximately the same mean of RP values. Conversely, in the rejection area, the SW test demonstrates slightly higher RP values than the LF test across different significance levels. For sample sizes of 10, 20 and 50, the AD test has the highest mean of RP values in the rejection area across different significance levels. In the non-rejection area, the SW test has the highest mean of RP values when α is 0.01 and 0.05. However, at $\alpha = 0.1$ both the SW and LF tests tend to have approximately the same mean of RP values, which are higher than those of the AD test. In the non-rejection area, the Shapiro-Wilk test tends to show the highest mean of RP values because it is designed to be sensitive to deviations from Normality, particularly in the central part of the distribution. Conversely, in the rejection area, the Anderson-Darling test showed the highest mean of RP values because it is more sensitive to deviation in the tail than the median, making it particularly effective in identifying extreme departures from Normality. The same results were obtained when performing simulations under H_1, if data are generated from $Exp(1)$, as shown in Appendix A.2.

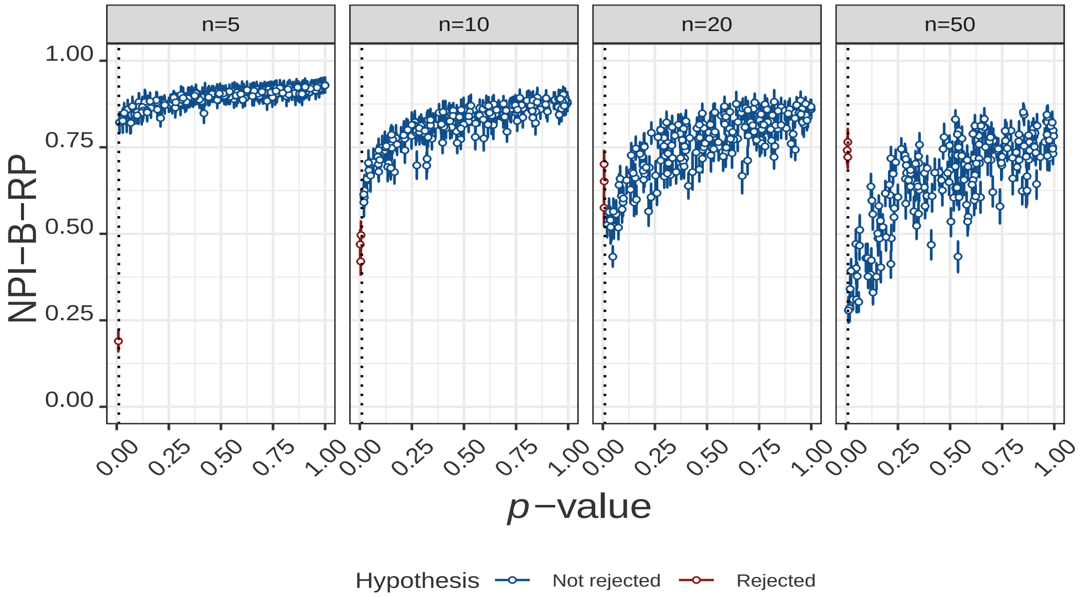


Figure 2.23: The relationship between NPI-B-RP and p-value for Shapiro-Wilk test for data sampled from $N(1, 1)$, $\alpha = 0.01$

In this chapter, we have explored the RP for Normality using various alternative hypothesis scenarios. To further validate and extend our findings, additional simulations were conducted where the data was generated from a mixture of Normal distributions. These simulations were designed to assess the RP under more complex distributional structures that might be encountered in practice.

Given the similarity in results to the previously discussed alternative examples, detailed outcomes of these additional simulations are included in Appendix A.1. These results reinforce the patterns observed with other alternative hypotheses, showing consistent RP values and providing further evidence for the robustness of our findings.

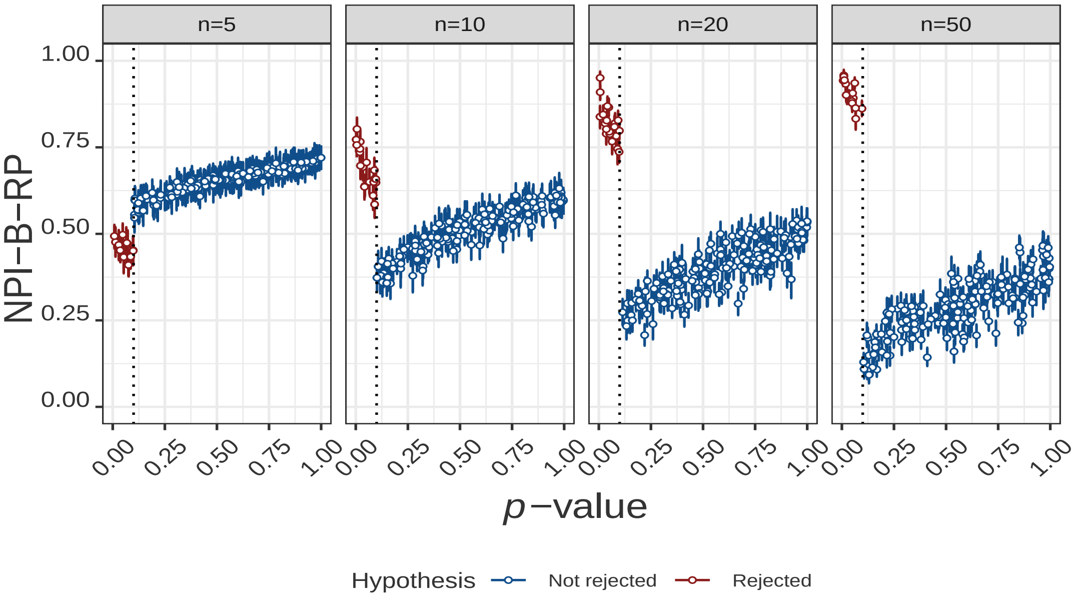


Figure 2.24: The relationship between NPI-B-RP and p-value for Shapiro-Wilk test for data sampled from $N(1,1)$, $\alpha = 0.1$

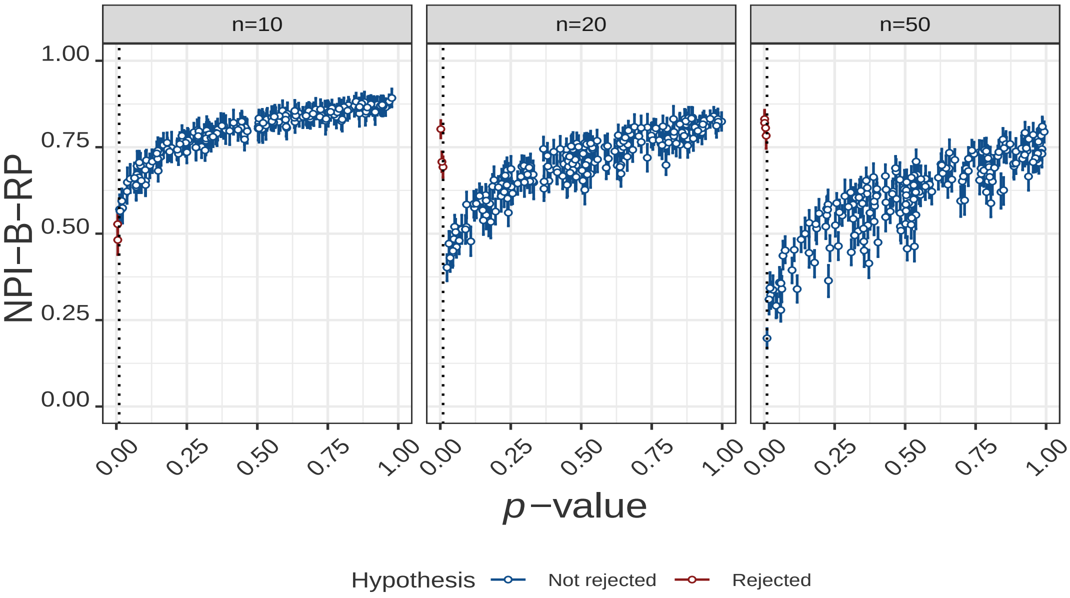


Figure 2.25: The relationship between NPI-B-RP and p-value for Anderson-Darling test for data sampled from $N(1,1)$, $\alpha = 0.01$

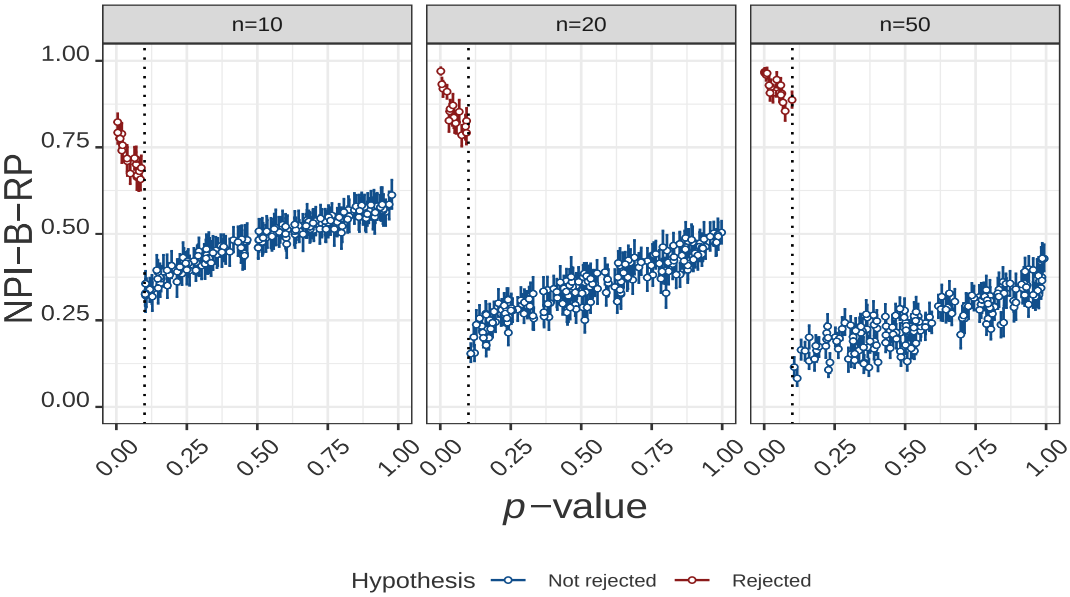


Figure 2.26: The relationship between NPI-B-RP and p-value for Anderson-Darling test for data sampled from $N(1,1)$, $\alpha = 0.1$

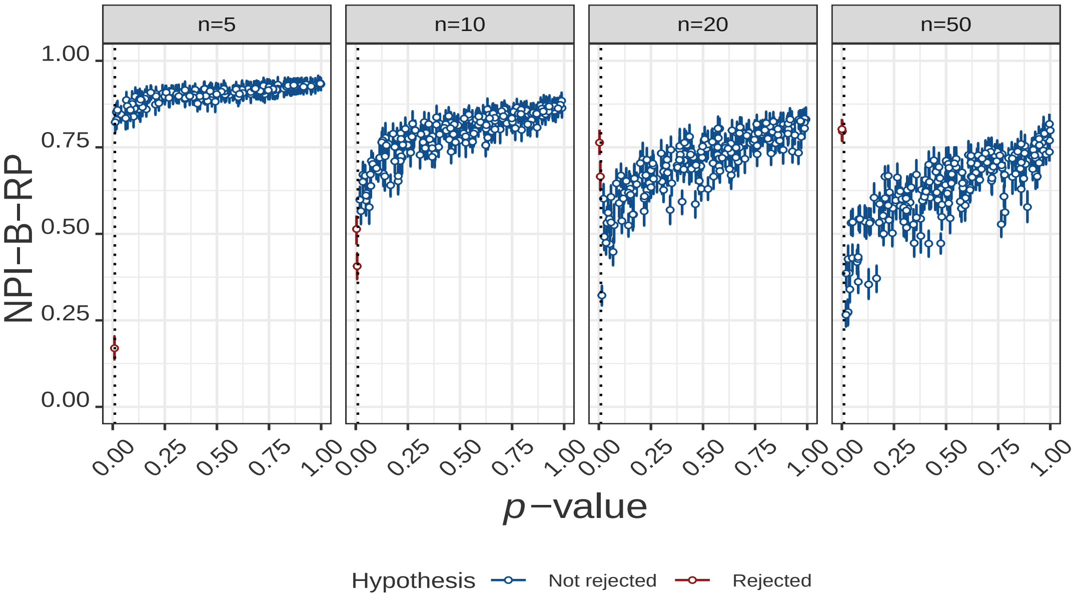


Figure 2.27: The relationship between NPI-B-RP and p-value for Lilliefors test for data sampled from $N(1,1)$, $\alpha = 0.01$

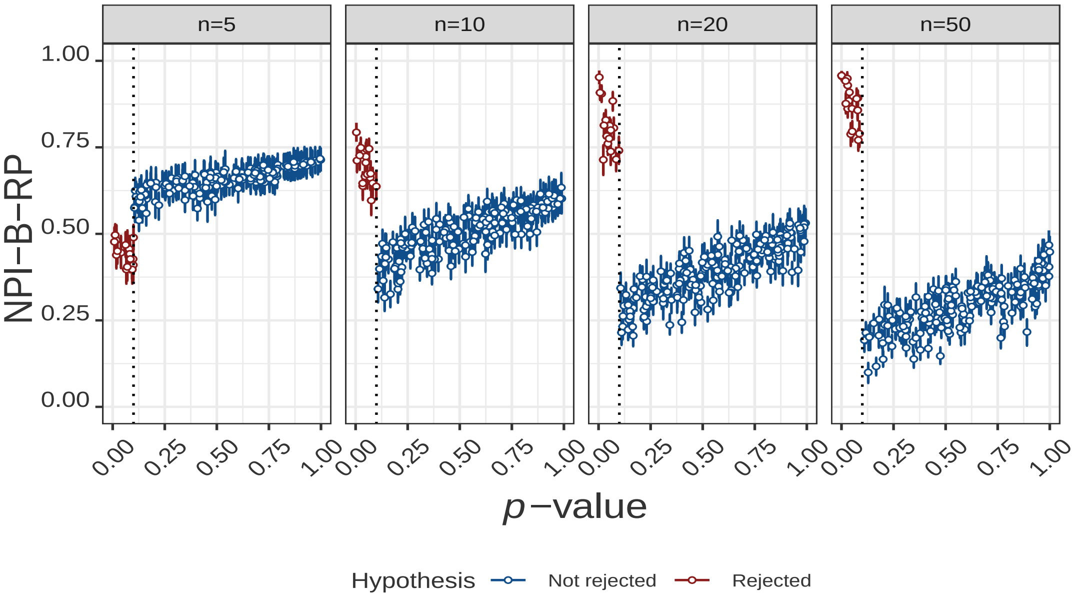
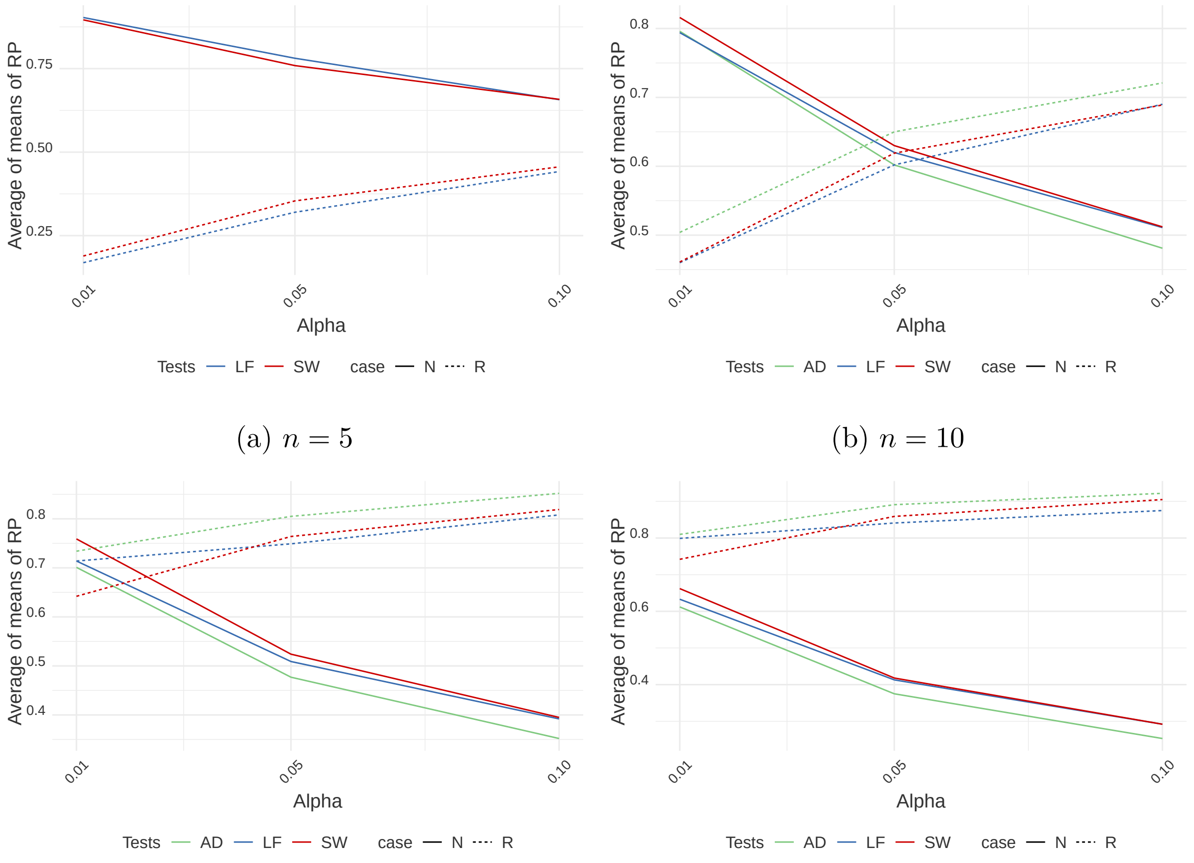


Figure 2.28: The relationship between NPI-B-RP and p-value for Lilliefors test for data sampled from $N(1, 1)$, $\alpha = 0.1$



(a) $n = 5$

(b) $n = 10$

(c) $n = 20$

(d) $n = 50$

Figure 2.29: The mean of RP values for Shapiro-Wilk, Anderson-Darling and Lilliefors tests for data sampled from $N(1, 1)$, for different levels of $\alpha = 0.01, 0.05, 0.1$ and with different sample sizes $n = 5, 10, 20, 50$.

2.4 Conclusions

In this chapter, the reproducibility (RP) for three Normality tests has been examined, namely the Shapiro-Wilk (SW) test, Anderson-Darling (AD) test and Lilliefors (LF) test. The RP values of these Normality tests were investigated through a simulation study under both the null hypothesis of Normality tests and their alternative hypothesis. These simulations contain various distributions, sample sizes, and significance levels.

The RP values for Normality tests tend to be low when the p-value of Normality tests is close to the level of significance, and RP values increase gradually as the p-value moves away from the level of significance. Generally, there is no substantial difference between the RP of these tests. However, the RP values of the AD test have less variability than RP values for the SW and LF tests. The AD test typically has the highest mean of RP values in the rejection area, whereas the SW test tends to have the highest mean of RP values in the non-rejection area.

The RP values are different according to sample size. For small sample sizes, RP values in the non-rejection area tend to be high, while RP values in the rejection area are low. Conversely, for large sample sizes, RP values in the non-rejection area are low and RP in the rejection area are high.

The relationship between the overall mean of RP values in the rejection area for the Normality tests with estimated power was examined. when the estimated power of the Normality tests increases, RP increases. Additionally, as the sample size increases, RP values in the rejection area and estimated power for the Normality tests increase. When data samples from a distribution that deviates much from Normal and for large sample sizes, the Normality tests tend to have approximately similar RP values in the rejection area, as well as similar estimated power. The Anderson-Darling test shows slightly high RP in the rejection area and estimated power when data are sampled from $Ca(0,1)$ distribution. The Shapiro-Wilk test has slightly high power but does not have high RP when data are drawn from $t(3)$ and $Exp(1)$ distributions. In most cases, the Lilliefors test displays the lowest power and RP.

The reproducibility for the Normality tests is investigated for different levels of significance, namely 0.01, 0.05 and 0.10. The observed pattern suggests that RP values in the non-rejection area are high and RP values in the rejection area are low when α is small. For larger values of α, RP values in the non-rejection decrease while RP values in the rejection area increase.

Chapter 3

Reproducibility of Tests for Equality of Variances

3.1 Introduction

The equality of variances assumption, also known as homoscedasticity or homogeneity of variances, is an important assumption in many statistical tests, particularly parametric statistical tests because they are sensitive to any difference. This assumption enables accurate inference and appropriate interpretation of test results, where unequal sample variances lead to biased and skewed test results. Common tests for the equality of variances are the F-test [93], Bartlett's test [10] and Leven's test [61]. Given the importance of the assumption of the equality of variances, the motivation of this chapter lies in investigating the reproducibility probability (RP) of tests of the equality of variances and understanding their role as preliminary tests in assessing the reproducibility of location tests in subsequent chapters.

The tests considered in this chapter are the F-test for equality of two variances and Levene's test. The reasons for choosing these tests are: F-test for equality of variances is a powerful test for detecting differences in variances between two samples, especially when the sample sizes are equal [18], and it is relatively simple to compute and interpret. Levene's test is less sensitive to departures from Normality than other equality of variances tests, making it a robust test [54].

This chapter briefly introduces the F-test and Levene's test for equality of variances in Section 3.2. To examine reproducibility probability (RP) for F-test and Levene's test and understand the relationship between their RP and p-values and the relationship between their RP and estimated power, a simulation study is conducted, and the investigation takes into account scenarios involving both null and alternative hypotheses in Section 3.3. Finally, the results and observations for RP for equality of variances tests are summarised in Section 3.5.

3.2 Tests for equality of variances

This section briefly introduces the F-test for equality of two variances and Levene's test, which can be applied to multiple variances.

3.2.1 F-test for equality of two variances

The F-test [93] is used to determine if there is a significant difference between the variances of the two samples. Suppose that there are two independent and identically distributed samples $\{X_1, X_2, \ldots, X_{n_X}\}$ and $\{Y_1, Y_2, \ldots, Y_{n_Y}\}$ from two populations that each have a Normal distribution, with sample sizes n_X and n_Y and variances S_X^2 and S_Y^2, respectively. The variances for the two populations σ_X^2 and σ_Y^2 are unknown, then the null hypothesis for F-test is expressed as $H_0 : \sigma_X^2 = \sigma_Y^2$ and the alternative hypotheses are for two-sided $H_1 : \sigma_X^2 \neq \sigma_Y^2$, for a lower one-tailed test $H_1 : \sigma_X^2 < \sigma_Y^2$, or for an upper one-tailed test $H_1 : \sigma_X^2 > \sigma_Y^2$. The test statistic for the F-test is [33]:

$$F = \frac{S_X^2}{S_Y^2} = \frac{\sum_{i=1}^{n_X}(X_i - \bar{X})^2/(n_X - 1)}{\sum_{i=1}^{n_Y}(Y_i - \bar{Y})^2/(n_Y - 1)} \tag{3.2.1}$$

where $\bar{X} = \frac{\sum X_i}{n_X}$ and $\bar{Y} = \frac{\sum Y_i}{n_Y}$ are the sample means.

For a two-sided test, the larger sample variance is placed in the numerator and the smaller sample variance in the denominator, ensuring the F-value is always greater than or equal to 1. The null hypothesis is rejected for a two-sided test if $F > F_{(\frac{\alpha}{2}, n_X-1, n_Y-1)}$ or $F < F_{(1-\frac{\alpha}{2}, n_X-1, n_Y-1)}$, for an upper one-tailed test if $F > F_{(\alpha, n_X-1, n_Y-1)}$, and for a lower one-tailed test if $F < F_{(1-\alpha, n_X-1, n_Y-1)}$, where $F_{(\alpha, n_X-1, n_Y-1)}$ is the critical value of the F-distribution with $n_X - 1$ degree of freedom of the numerator and $n_Y - 1$ denominator degrees of freedom, and α is the significance level. F-test is highly sensitive to departures from Normality and to outliers [77].

3.2.2 Levene's test for equality of variances

Levene's test [61] serves to examine the equality of variances among multiple groups or samples. The assumption of equal variances across groups is an important assumption for some statistical tests, such as analysis of variance (ANOVA) [73, 100].

Assume that samples $\{X_{ij} : j = 1 \ldots, n_i, \, i = 1, \ldots, M\}$ from M populations with mean μ_i and variance σ_i^2 for the i-th population, and distribution function $F\left(\frac{X-\mu_i}{\sigma_i}\right)$. The function F and the constants μ_i, and σ_i are unknown [63]. The null hypothesis is $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_M^2$, against the alternative hypothesis $H_1 : \sigma_i^2 \neq \sigma_j^2$ not all the population variances are equal [63].

The test statistic of Levene's test is based on a one-way analysis of variance (one-way ANOVA) using the values $Z_{ij} = |X_{ij} - \tilde{X}_i|$, where $\tilde{X}_i$ is the mean (or median) of the i-th population [55, 63]. The test statistic is:

$$L = \frac{(n_T - M)}{(M-1)} \cdot \frac{\sum_{i=1}^{M} n_i (\bar{Z}_{i.} - \bar{Z}_{..})^2}{\sum_{i=1}^{M} \sum_{j=1}^{n_i} (Z_{ij} - \bar{Z}_{i.})^2} \tag{3.2.2}$$

where

$$\bar{Z}_{i.} = \frac{\sum_{j=1}^{n_i} Z_{ij}}{n_i}, \quad \text{and} \quad \bar{Z}_{..} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{n_i} Z_{ij}}{n_T}$$

where n_T is the total number of observations in all groups.

When L exceeds the $100(1-\alpha)$th percentile of the F-distribution with degrees of freedom $(M-1)$ and $(n_T - M)$, the null hypothesis is rejected [63]. Levene's test does not require Normality of the underlying data [96]. Levene's test is built upon the F-distribution for the test statistic, which typically does not differentiate between the directions of the alternative hypothesis (i.e., whether the variances are greater or less than each other). Therefore, this thesis is limited to studying the reproducibility of the standard Levene's test for the two-sided alternative hypothesis.

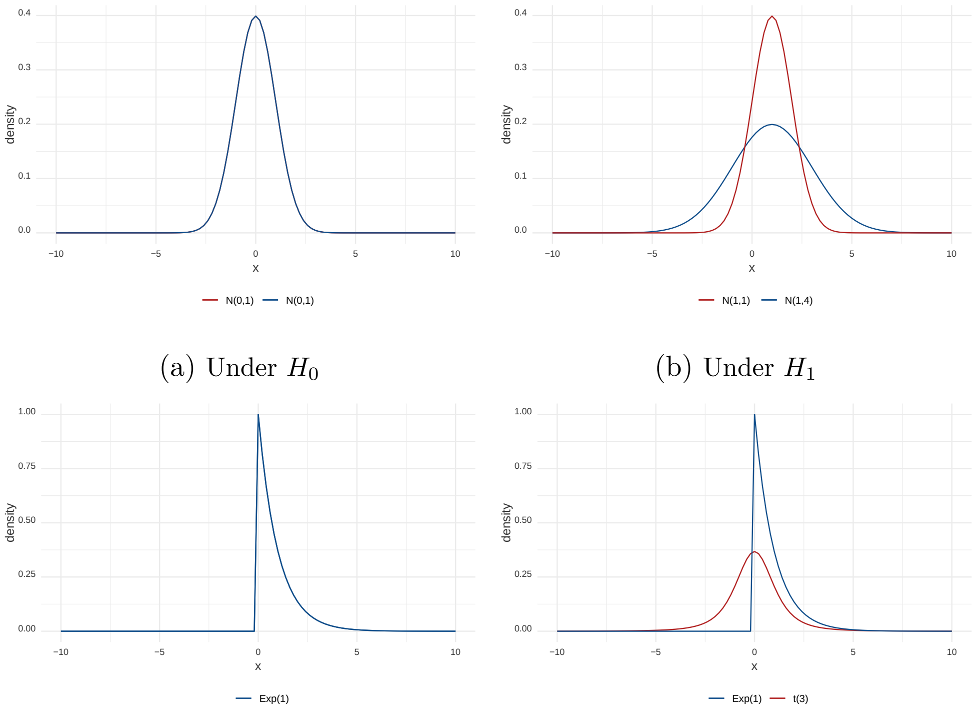
3.3 Simulation studies for reproducibility of the tests for equality of variances

In this section, we explore the reproducibility of F-test and Levene's test using simulation studies. The reproducibility is computed using the NPI-B-RP Algorithm 1, as presented in Section 1.4.5 of Chapter 1. The inputs are the two original samples with the sample sizes n_X and n_Y, respectively.

The simulations cover scenarios where the alternative hypothesis is two-sided for both tests. Specifically, we simulate data under the null hypothesis $H_0 : \sigma_i^2 = \sigma_j^2$ and under the alternative hypothesis $H_1 : \sigma_i^2 \neq \sigma_j^2$, where $i \neq j$. Additionally, we extend the investigation to include scenarios in which the alternative hypothesis for the F-test is upper one-tailed, denoted as $H_1 : \sigma_X^2 > \sigma_Y^2$.

Data are generated from various distributions to conduct the simulations, and their probability density functions (PDFs) are shown in Figure 3.1. Under the null hypothesis H_0, data are generated from Normal distributions with a mean of 1 and a standard deviation of 1 for both samples, denoted by $N(1, 1)$, therefore $\sigma_X^2 = \sigma_Y^2 = 1$. Additionally, data are generated from a non-Normal distribution, specifically the Exponential distribution with a rate of 1 for both samples, denoted by $Exp(1)$, with $\sigma_X^2 = \sigma_Y^2 = 1$. Under the alternative hypothesis H_1, data are sampled from Normal distributions $N(1, 2^2)$ and $N(1, 1^2)$, with $\sigma_X^2 = 4$ and $\sigma_Y^2 = 1$. Data are also drawn from non-Normal distributions $t(3)$ and $Exp(1)$, with $\sigma_X^2 = 3$ and $\sigma_Y^2 = 1$. The reason for choosing these distributions is that the Normal distributions are chosen to examine RP for the F-test and Levene's test when the Normality assumption is met, both for samples with the same variances and different variances. Secondly, the non-Normal distributions are chosen to examine the effects of violating the Normality assumption on RP for both the F-test and Levene's test, again for samples with the same variances and different variances.

The number of runs per simulation is $K = 200$, for each run, two original samples of the same sample size n are generated from the chosen distributions, an equality of variances test is performed on these original samples and NPI-B-RP is computed using



(a) Under H_0

(b) Under H_1

(c) Under H_0

(d) Under H_1

Figure 3.1: PDFs for the investigated distributions.

the NPI-B-RP Algorithm 1. Data are simulated for sample sizes $n_X = n_Y = 10, 25$. The tests are performed with 5% level of significance.

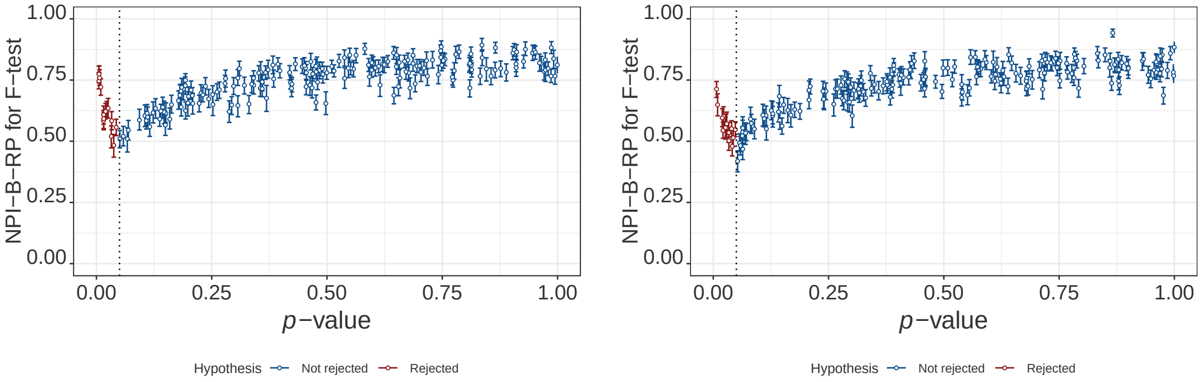
3.3.1 Simulation results for the reproducibility of F-test

This subsection presents the simulation results for estimating the NPI-B-RP for F-test for the equality of two variances. From Figure 3.2, which presents RP values for the F-test in the case of the two-sided test, under the null hypothesis H_0 when both original samples are drawn from Normal distribution with $\sigma_1^2 = \sigma_2^2 = 1$. The RP values demonstrate a general pattern: RP tends to be low when the p-values for the F-test are close to the significance level $\alpha = 0.05$. Conversely, when the p-value moves away from this threshold, RP values increase in both the rejection and non-rejection areas. Furthermore, the RP does not converge to one when the p-value approaches one. This is because when the p-value is close to one in a two-sided test, it suggests limited evidence against H_0 and does not offer certainty about the actual equality of variances. This uncertainty is reflected in the RP, which may not converge to one as the p-value approaches one.

Figure 3.3 shows RP values for F-test in the case of the two-sided test, under H_1 when both original samples are drawn from Normal distribution with different variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 1$. The relationship between the RP values and the p-values shows a similar relationship to that observed when simulating under H_0, with the difference being that most of the original samples are in the rejection area with high RP values. For example, with a sample size of 10, there are 104 original samples in the rejection area, whereas for a sample size of 25, there are 181 samples.

The power for the F-test is estimated by conducting a Monte Carlo simulation for $10,000$ datasets that are simulated from Normal distributions with different variances $N(1, 2^2)$ and $N(1, 1^2)$ respectively. For each simulated dataset, we perform the F-test for equality of two variances. Then determine how often the test correctly rejects H_0. The proportion of times H_0 is rejected out of the total number of simulations provides an estimate of the power of the F-test. The relationship between the overall mean of RP values in the rejection area and estimated power is examined. For sample size 10, the estimated power equals 0.4943 whereas the mean of RP values in the rejection area is equal to 0.685. For the sample size of 25, the power is equal to 0.917 and the mean of RP values is 0.797. It is clear that as the power increases, the mean of RP values increases in the rejection area.

Figure 3.4 illustrates the outcomes of RP for the F-test when both original samples are drawn from $Exp(1)$ under H_0 ($\sigma_1^2 = \sigma_2^2 = 1$) for the two-sided hypothesis. RP values show the general pattern, with a notable observation that, although data is being simulated under H_0, there is a large number of original samples located in the rejection area with RP values exceeding 50%, in contrast to when samples are drawn from the Normal distribution. Moreover, RP values show greater variability than those observed when samples are drawn from the Normal distribution, especially in the non-rejection area. This is due to the nature of data distributions and sensitivity of F-test to deviation from Normality; original samples taken from $Exp(1)$ deviate substantially from Normality and generate NPI-B samples that also tend to be skewed. Consequently, the Type I error rate will be much higher than α due to the violation of Normality, leading to increased variability in RP values and the number of original samples that located in the rejection area.



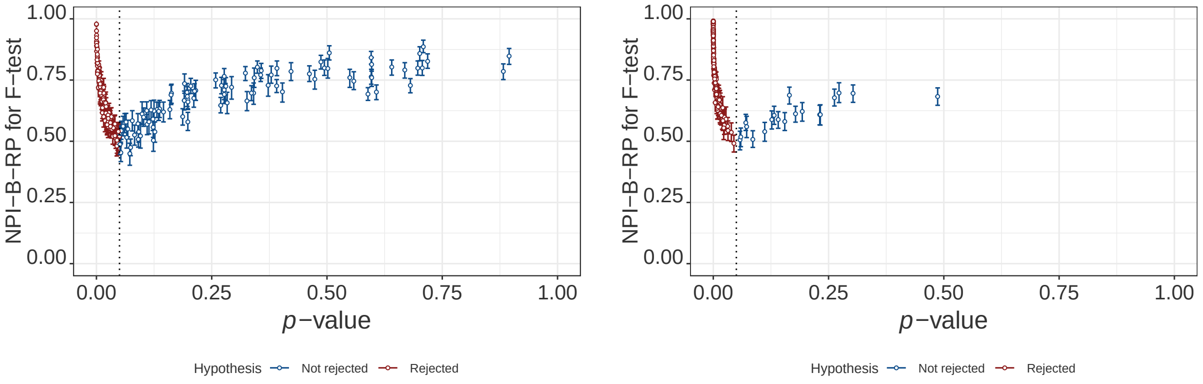
(a) $n = 10$

(b) $n = 25$

Figure 3.2: The relationship between p-values and NPI-B-RP for the F-test, under H_0, both samples sampled from $N(1, 1^2)$, in a two-sided test

Figure 3.5 presents RP values for the F-test when original samples are drawn from $t(3)$ and $Exp(1)$ under H_1 ($\sigma_1^2 = 3$, $\sigma_2^2 = 1$) for the two-sided hypothesis. RP values show the general pattern, the most original samples are located in the rejection area. The number of samples in the rejection area increases with increasing the sample size, for example, there are 70 samples in the rejection area for a sample size of 10, whereas there are 114 samples for a sample size of 25. There is variability in RP values in the non-rejection area, but it appears to be diminished in the rejection area. This could be because, in the non-rejection area, there is a wider range of possible data distributions with varying degrees of variance inequality can still lead to non-rejection of H_0. This leads to greater variability in the RP values. However, where H_0 is rejected, the results are more constrained in the rejection area, resulting in reduced variability in RP values. The relationship between the overall mean of RP values in the rejection area and the estimated power for the F-test is examined. For the sample size of 10, the power of the F-test is equal to 0.357 and the mean of RP in the rejection area is 0.739. For the sample size of 25, the power of the F-test is equal to 0.555 and the mean of RP in the rejection area is 0.768. Thus, it is clear that as the power of the F-test increases RP for the F-test in the rejection area increases.

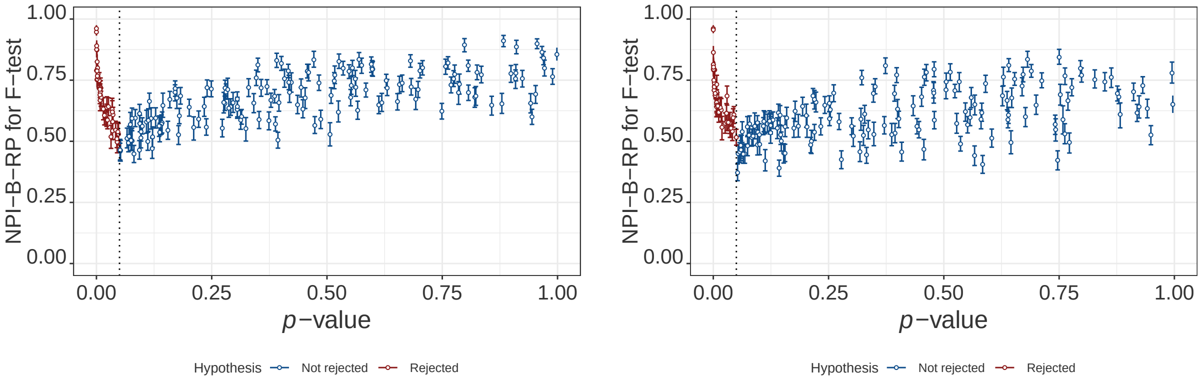
For the upper one-tailed F-test, Figure 3.6 shows RP values for F-test when data are simulated under H_0 and both original samples are drawn from $N(1, 1^2)$. The RP values



(a) $n = 10$

(b) $n = 25$

Figure 3.3: The relationship between p-values and NPI-B-RP for the F-test, under H_1, samples sampled from $N(1, 2^2)$ and $N(1, 1^2)$, in a two-sided test

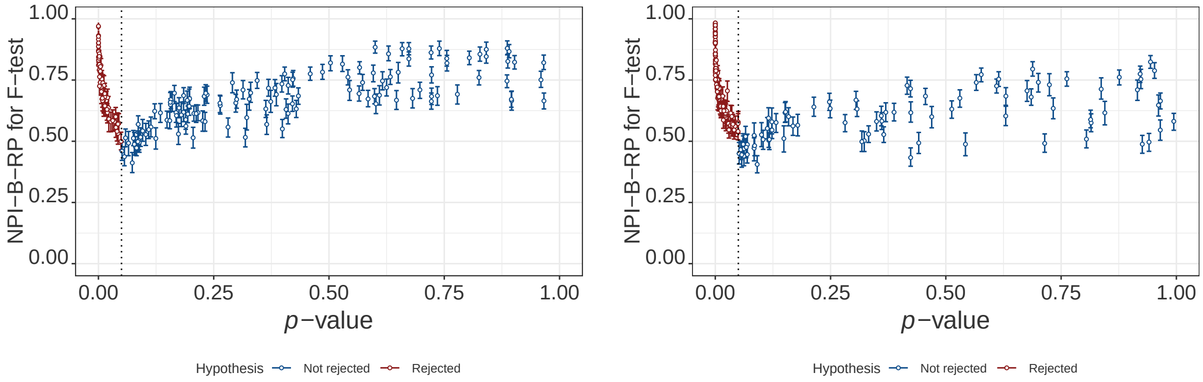


(a) $n = 10$

(b) $n = 25$

Figure 3.4: The relationship between p-values and NPI-B-RP for the F-test, samples sampled from $Exp(1)$ under H_0, in a two-sided test

for the F-test show the general pattern and they show less variability compared to the case of the two-sided test. The decreased variability in RP values for the one-sided upper test, as opposed to the two-sided test, can be attributed to the nature of the examined alternative hypothesis. For the upper one-tailed F-test is designed to identify whether the variance of one population is significantly greater than the variance of another population without considering whether the second might be larger. This hypothesis limits the range of potential results because it is more effective at detecting such differences when they exist, leading to reduced variability in test outcomes. Conversely, the two-sided test considers both possibilities of unequal variances in either direction, resulting in a larger



(a) $n = 10$

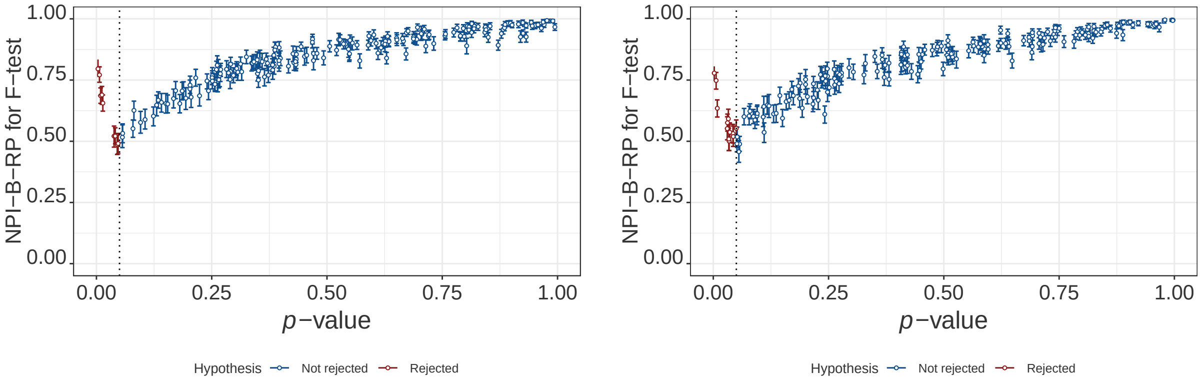
(b) $n = 25$

Figure 3.5: The relationship between p-values and NPI-B-RP for the F-test, under H_1, samples sampled from $t(3)$ and $Exp(1)$, in a two-sided test

range of potential outcomes and therefore increased variability in RP values. For the same reason, RP values for the upper one-tailed F-test go close to one when the p-values are close to one.

Similarly, when simulate data under H_1 with samples from Normal distributions with different variances $\sigma_1^2 = 4$ and $\sigma_2^2 = 1$, the RP values are shown in Figure 3.7. RP values tend to be close to one as the p-value decreases close to zero. Moreover, for sample size 10, there are 125 original samples in the rejection area of 200 samples and their mean of RP values is equal to 0.705 whereas the estimated power equals 0.628. For the sample size 25, the power of the test is equal to 0.950 and the mean of RP values in the rejection area is equal to 0.853 for 192 original samples in the rejection area. Thus, as the power increases, the mean of RP in the rejection area for the upper one-tailed F-test also increases. Additionally, as the sample size increases, both the mean of RP in the rejection area and the power increase.

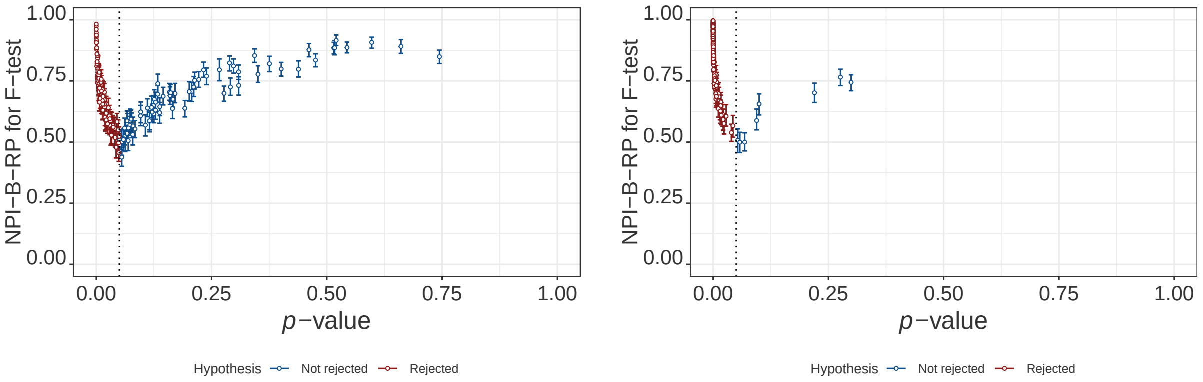
Figure 3.8 illustrates the outcomes of RP for the F-test when both original samples are drawn from $Exp(1)$ under H_0 for the upper one-tailed hypothesis. RP values show a pattern similar to that observed when data are drawn from Normal distributions as shown in Figure 3.6, with slight variability in RP values in the non-rejection area. Likewise Figure 3.9 that shows RP values for upper one-sided F-test when data are drawn from non-Normal distributions with different variances $\sigma_1^2 = 3$ and $\sigma_2^2 = 1$. Most original



(a) $n = 10$

(b) $n = 25$

Figure 3.6: The relationship between p-values and NPI-B-RP for the F-test, under H_0, samples sampled from $N(1, 1^2)$, in an upper one-sided test

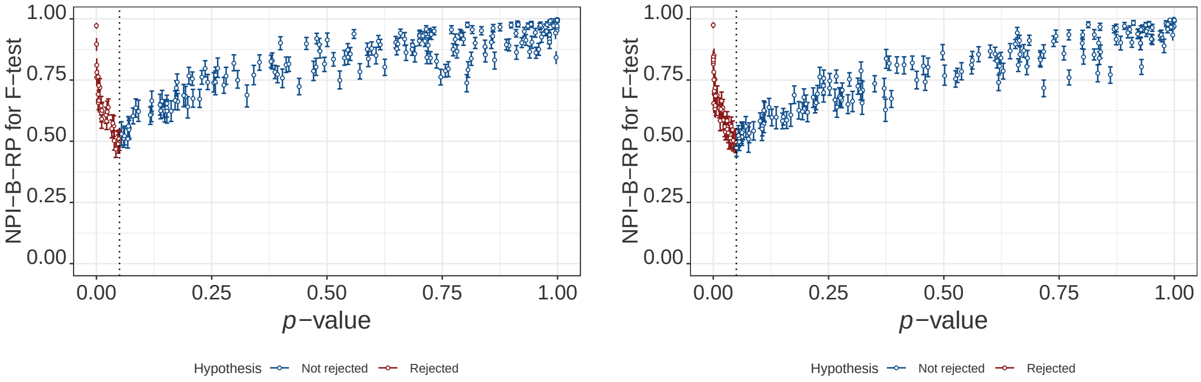


(a) $n = 10$

(b) $n = 25$

Figure 3.7: The relationship between p-values and NPI-B-RP for the F-test, under H_1, samples sampled from $N(1, 2^2)$ and $N(1, 1^2)$, in an upper one-sided test

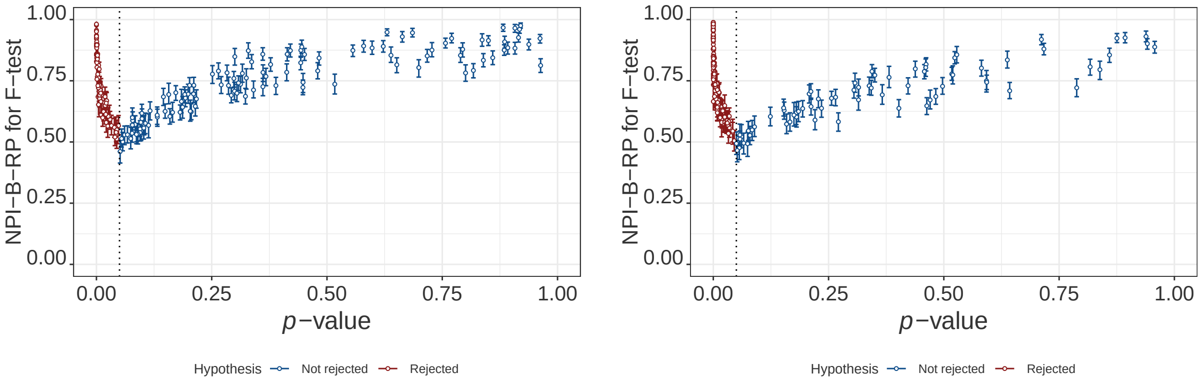
samples are located in the rejection area, and the number decreases in the non-rejection area as the sample size increases. For a sample size of 10, there are 83 samples in the rejection area, whereas for a sample size of 25, there are 128 samples. The mean of RP values for the F-test in the rejection area increases as the estimated power for the F-test increases. For sample size 10, the power is equal to 0.414 and the mean of RP values in the rejection area is equal to 0.745. For sample size 25, the power is equal to 0.615 and the mean of RP values in the rejection area is equal to 0.772.



(a) $n = 10$

(b) $n = 25$

Figure 3.8: The relationship between p-values and NPI-B-RP for the F-test, samples sampled from $Exp(1)$ under H_0, in the upper one-tailed test



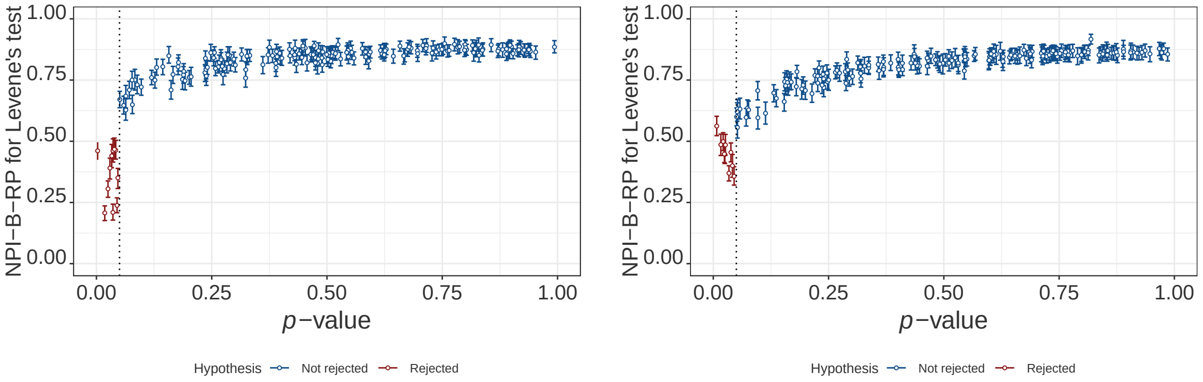
(a) $n = 10$

(b) $n = 25$

Figure 3.9: The relationship between p-values and NPI-B-RP for the F-test, under H_1, samples sampled from $t(3)$ and $Exp(1)$, in an upper one-sided test

3.3.2 Simulation results for the reproducibility of Levene's test

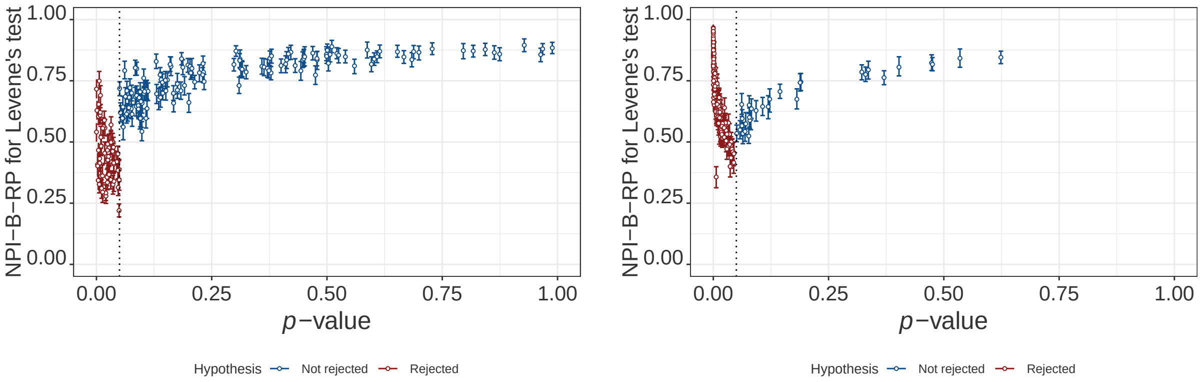
This section shows the results of the simulation study that estimates the NPI-B-RP for Levene's test for equality of variances. Figure 3.10 shows the relationship between p-values and RP values for Levene's test, when both original samples are drawn from Normal distribution with the same variances $\sigma_1^2 = \sigma_2^2 = 1$. This relationship shows the general pattern: RP is low when p-value is close to the level of significance, and RP tend to be high as p-value is far away from the level of significance. In the rejection area, where the null hypothesis is rejected with a given significance level of 0.05, the RP values tend



(a) $n = 10$

(b) $n = 25$

Figure 3.10: The relationship between p-values and NPI-B-RP for Levene's test, under H_0, samples from $N(1, 1^2)$, in a two-sided test



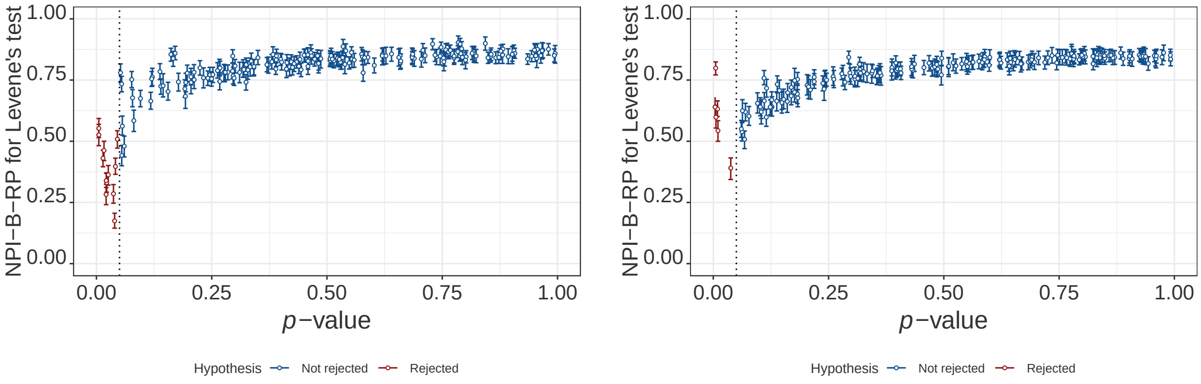
(a) $n = 10$

(b) $n = 25$

Figure 3.11: The relationship between p-values and NPI-B-RP for Levene's test, under H_1, samples from $N(1, 2^2)$ and $N(1, 1^2)$, in the two-sided test.

to be less than 50%, it means that the results are not very reproducible.

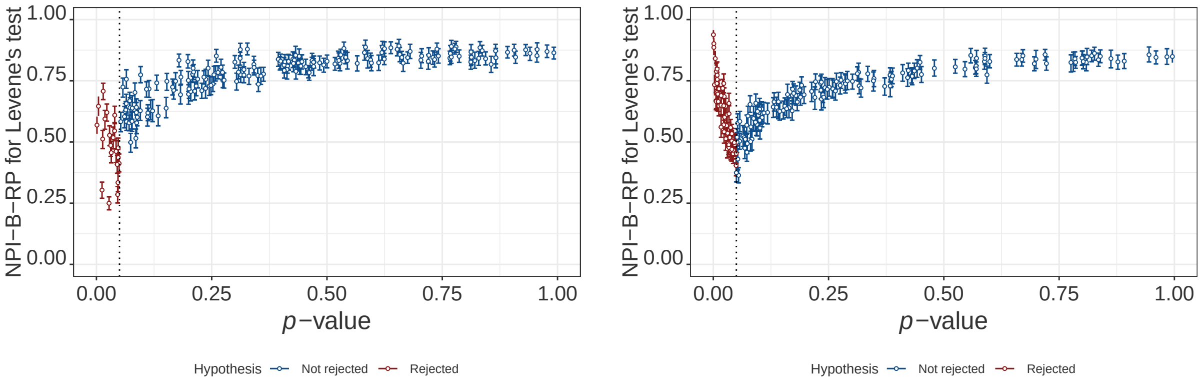
Figure 3.11 shows RP values for Levene's test when data are drawn from Normal distributions with different variance $\sigma_1^2 = 4$ and $\sigma_1^2 = 1$. For sample size 10, RP has strong variability, especially when the p-value is close to the threshold. For the sample size of 25, RP has no noticeable variability and most original samples are located in the rejection area. Moreover, as the estimated power for Levene's test and the sample size increase, the RP for Levene's test in the rejection area increases. To clear that for sample size 10, the estimated power of Levene's test equals 0.497 and the mean of RP values in



(a) $n = 10$

(b) $n = 25$

Figure 3.12: The relationship between p-values and NPI-B-RP for Levene's test, samples from $Exp(1)$ under H_0, in the two-sided test



(a) $n = 10$

(b) $n = 25$

Figure 3.13: The relationship between p-values and NPI-B-RP for Levene's test, samples from $t(3)$ and $Exp(1)$ under H_1, in the two-sided test

the rejection area is equal to 0.453, whereas for sample size 25, the power is equal to 0.914 and RP is equal to 0.700.

Figure 3.12 shows RP values for Levene's test if both original samples are drawn from non-Normal distributions with the same variances $\sigma_1^2 = \sigma_2^2 = 1$. RP show the general pattern, RP values in the rejection area for sample size 10 seem low reaching 0.5, while in the non-rejection area are high. For sample size 25, RP values in the rejection area are higher than those for sample size 10. Similarly, RP values for Levene's test if original samples are drawn from non-Normal distributions with different variances $\sigma_1^2 = 3$ and

$\sigma_2^2 = 1$ that are shown in Figure 3.13. For the sample size of 10, there is variability in RP values that are close to the threshold.

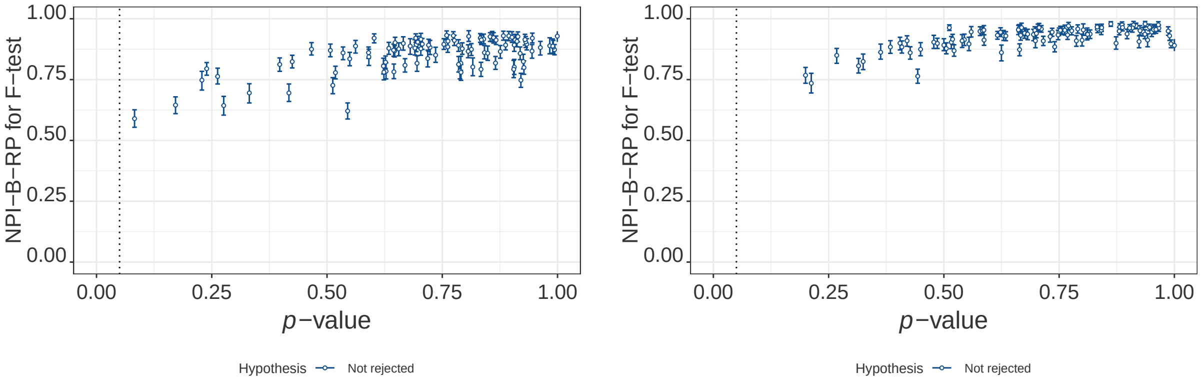
In later chapters of the thesis, Levene's test is utilized as the preliminary test for scenarios involving more than two groups. Therefore, we examine RP for Levene's test for three original samples. The results are approximately similar to RP for the two samples but RP in the non-rejection area tends to be slightly lower than RP for two samples, and RP in the rejection area tend to be slightly higher than RP for two samples. The results are presented in Appendix B.

Additional simulations were conducted with data generated from a mixture of Normal distributions under H_0 to further explore the reproducibility probability (RP) in the context of the F-test for equality of two variances and Levene's test. Specifically, the data for each group were generated from a mixture model defined as:$X \sim 0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$

In these simulations, it was observed that the RP is high in the non-rejection area as shown in Figures 3.14 and 3.15, indicating a strong probability that repeated tests will yield consistent results. Additionally, it was noted that when performing F-test all original samples fell into the non-rejection area, with p-values slightly far away from the significance threshold as shown in Figure 3.14. This suggests that the variances of the mixture Normal distributions do not differ enough to reject the null hypothesis under the F-test. For Levene's test, there are two original samples in the rejection area with very small RP values that do not exceed 0.50, indicating non-reproducibility, as shown in Figure 3.15.

3.4 Comparison between reproducibility of F-test and Levene's test

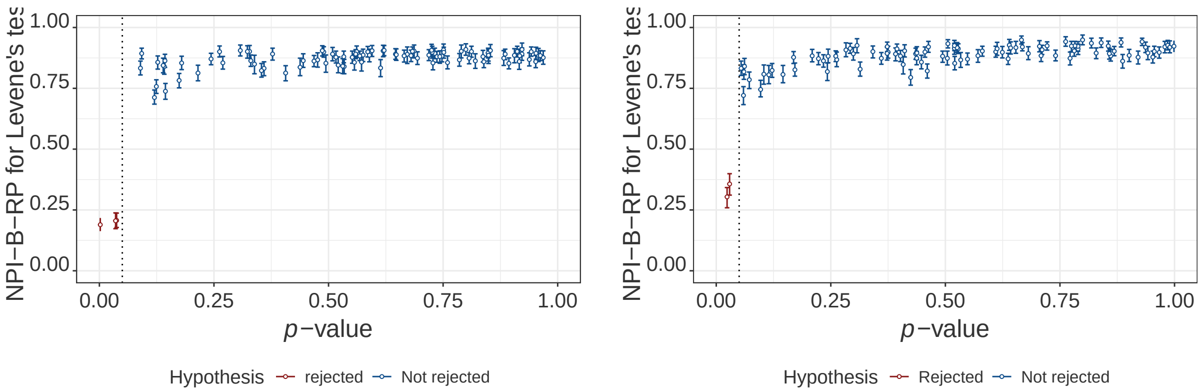
In this part, we discuss the differences between reproducibility for F-test and Levene's test. For Levene's test, RP values show less variability than RP for the two-sided F-test, particularly when dealing with non-Normal data. This is because of the diversity in the NPI-B samples distributions, as well as the sensitivity of the F-test to deviations from



(a) $n = 10$

(b) $n = 25$

Figure 3.14: The relationship between p-values and NPI-B-RP for F-test, samples from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, under H_0, in the two-sided test



(a) $n = 10$

(b) $n = 25$

Figure 3.15: The relationship between p-values and NPI-B-RP for Levene's test, samples from a mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, under H_0, in the two-sided test

Normality and outliers. In contrast, Levene's test does not require the assumption of Normality. This difference in sensitivity to Normality assumptions between the two tests could explain the lower variability in RP values observed in Levene's test. Moreover, in the rejection area, RP values for Levene's test and the number of original samples tend to be smaller than those for the F-test and their RP value does not exceed 0.5 when simulating data under H_0. Furthermore, the effect of sample size on RP values differ between the F-test and Levene's test when dealing with non-Normal data. Larger sample sizes result in more stable RP values for Levene's test, conversely, RP values for the F test show greater variability, especially in the non-rejection area.

3.5 Conclusions

In conclusion, this chapter focused on investigating reproducibility probability (RP) for the tests examining the equality of variances, specifically the F-test and Levene's test. This investigation is an integral part of the broader exploration of this thesis which focuses on reproducibility for multi-stage procedures, particularly in the context of location tests which require investigation from the assumption of equality of variances through preliminary tests such as F-test and Levene's test. Through simulation studies, scenarios involving the two-sided hypothesis and one-sided hypotheses under H_0 and H_1 for the F-test are examined. Additionally, the RP for Levene's test of the two-sided hypothesis under H_0 and H_1 are studied.

The results for both tests revealed that the RP value show the general pattern for RP: RP is low when the p-value is close to the threshold, and increases as the p-value moves away from this threshold. Furthermore, for two-sided hypotheses, RP for both tests does not approach one when the p-value is close to one. Whereas for an upper one-sided F-test, RP tends to approach one as the p-value approaches one. Additionally, it is noticed that the case of two-sided RP for the F-test has variability more than the upper one-tailed F-test. When comparing RP for Levene's and F-test, the RP for Levene's test has less variability than RP for the F-test especially when dealing with non-Normal data.

For the relationship between the overall mean of RP values in the rejection area and the estimated power of the equality of variances tests, it is clear that as the power of the test increases, the mean of RP values in the rejection area also increases. Additionally, regarding the relationship between the sample size, power, and RP in the rejection area, as the sample size increases, both the power and RP values increase.

It can be said that The F-test's sensitivity to departures from the Normal assumption contributes to greater variability in RP values. While, the RP values of Levene's test perform better than those for the F-test, depending on less variability in RP values and low RP values in the rejection area when drawing from a Normal distribution. Therefore, it is important to take into account the underlying assumptions of each test when choosing a suitable approach for evaluating equality in variance. In cases where the data are known to deviate substantially from Normality or contain outliers, opting for Levene's test may be more robust and better reproducibility as it does not depend on the assumption of Normality.

Chapter 4

Reproducibility of One-Sample Location Tests with and without Preliminary Test

4.1 Introduction

In statistics, the inference of the population mean is one of the most fundamental concepts, which is typically done using location tests. One-sample location tests are a powerful statistical tool for assessing whether a sample mean is consistent with a hypothesized value or standard, and can be applied to a wide range of real-life scenarios. The results of these tests can be used to inform decision-making processes in various fields, including research, business, and healthcare. For example, in the field of healthcare, the one-sample t-test is commonly employed to assess the efficacy of new treatments or medical equipment.

In this chapter, we focus on two types of one-sample location tests: the one-sample t-test and the one-sample Wilcoxon signed-rank test. The t-test is a parametric test that requires the assumption of Normality. The Wilcoxon test is a nonparametric test that does not depend on any specific distributions.

The one-sample t-test is the most commonly used parametric location test to determine whether a population mean is significantly different from a hypothesized value [65]. Suppose that the hypothesized value for the population mean is μ_0, and μ is the unknown population mean, the null hypothesis for t-test is $H_0 : \mu = \mu_0$, against the alternative hypothesis $H_1 : \mu \neq \mu_0$ [39]. The t-statistic follows a Student's t-distribution with $n - 1$ degrees of freedom [39, 72]:

$$t = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \tag{4.1.1}$$

where $\bar{X}$ is the sample mean, n is the sample size, and S is the sample standard deviation.

A commonly used nonparametric test is the one-sample Wilcoxon signed-rank test. The null hypothesis for the one-sample Wilcoxon signed-rank test is $H_0 : \eta = \eta_0$ against the alternative hypothesis $H_1 : \eta \neq \eta_0$, where η is the population median and η_0 is the hypothesized value of the median in the population [103]. Suppose that $X_1, X_2, \ldots, X_n$ is a random sample from the distribution of the continuous random variable X. The test statistic for the one-sample Wilcoxon signed-rank test W is [43]

$$W = \min(W_+, W_-) \tag{4.1.2}$$

where

$$W_+ = \sum_{i=1}^{n} R_i \quad \text{for all positive ranks} \tag{4.1.3}$$

$$W_- = |\sum_{i=1}^{n} R_i| \quad \text{for all negative ranks} \tag{4.1.4}$$

where R_i is the rank of the difference $|X_i - \eta_0|$, for $i = 1, \ldots, n$.

In this chapter, the reproducibility probability (RP) for a two-stage procedure is investigated. The two-stage procedure involves a preliminary test for Normality followed by either the one-sample t-test or the Wilcoxon test, depending on the outcome of the Normality test. Additionally, the reproducibility of one-sample location tests without conducting a preliminary test for Normality is explored. We aim to assess the impact of a preliminary Normality test on the RP of the one-sample location tests by comparing the RP values of location tests obtained with and without preliminary test. Furthermore, the relationship between RP values for one-sample location tests and their estimated power is examined.

Section 4.2 addresses the reproducibility probability (RP) for the two-stage procedure (RP for one-sample location tests with a preliminary test for Normality). In Section 4.3, simulation studies for the reproducibility of the two-stage procedure and the reproducibility of the one-sample location tests without the preliminary test for Normality are performed, this section also presents the simulation studies' results. The impact of the preliminary Normality test on RP for location tests is addressed in Section 4.4. Finally, Section 4.5 provides a summary of the study.

4.2 Reproducibility for location tests with the preliminary test of Normality

This section aims to evaluate the reproducibility of the two-stage procedure. In this two-stage procedure, the initial stage involves assessing the assumption of Normality using a preliminary test of Normality. If the null hypothesis for Normality at the first stage is not rejected, then the analysis proceeds with the use of the one-sample t-test. If the null hypothesis for Normality is rejected at the first stage, then the one-sample Wilcoxon test is employed.

The reproducibility of the two-stage procedure can be assessed in various ways such as:

1. *Case A*: Full reproducibility for the two-stage procedure, that is both the preliminary test and the location test lead to the same conclusion.

2. *Case B*: Reproducibility for the same outcome for the location test, no matter which test is used.

3. *Case C*: Reproducibility of the location test conclusion, where for the bootstrap samples the same location test is applied as for the original sample without further preliminary testing.

The motivations and aims for studying these cases are:

Case A: The aim is to examine if the combined two-stage RP is noticeably different from the product of RPs for the two individual tests (preliminary test and location test). The motivation behind this case study is to investigate the impact of a preliminary test on the reproducibility of location tests. By contrasting the combined RP with the product of individual RPs, we can evaluate whether the inclusion of a preliminary test enhances or diminishes the reproducibility of location tests.

Case B: The aim is to consider whether the application of preliminary tests to choose the appropriate location tests enhances the reproducibility of the outcome of the location test. The motivation behind this case is to understand the impact of performing a preliminary test on the reproducibility of the outcome of the location test.

This case may be important from a practical perspective. For example, if testing whether a new medication has a noticeable impact on the treatment of a disease. The two-stage procedure is applied, the initial stage involves conducting a preliminary test to investigate Normality. Subsequently, in the second stage, either the one-sample t-test is conducted if the data follow a Normal distribution or the one-sample Wilcoxon signed-rank test is used if the data do not follow a Normal distribution. Here, the reference value is zero. if the location parameter (mean or median) is either smaller or larger than zero, it indicates the effect of the new medication. Our interest is in the reproducibility of the outcome (rejection or non-rejection) of the null hypothesis for location tests whether we get this via the t-test or the Wilcoxon test. This two-stage procedure can provide insights into the reproducibility of the location test outcomes (the effect of medication or not). By comparing the RP values for this case with those for the location test without the preliminary test, the researcher can assess the impact of the preliminary test on the reproducibility of the results.

Case C: This case aims to investigate whether filtering original samples based on a preliminary test's outcome and applying a location test to all NPI-B samples without further preliminary testing can improve the reproducibility of the original results, compared to performing a location test without preliminary test. The motivation behind *Case C* is to assess whether preliminary test filtering can help identify cases where the original samples exhibit certain characteristics, such as increased skewness, which lead to NPI-B samples also tending to be skewed. This skewness could negatively (or positively) impact the RP of the location test. The goal is to determine if this approach can lead to a higher RP compared to simply applying the location test without any pre-filtering based on the preliminary test.

The following steps assess the reproducibility probability values with the NPI-B samples for the two-stage procedure. Suppose that N is the number of NPI-B samples and N_t^* and N_W^* are two disjoint subsets of $\{1, 2, \dots, N\}$, $N_t^* \cap N_W^* = \emptyset$, where N_t^* represents all the indices for the NPI-B samples that pass the Normality test and perform the t-test, and N_W^* represents all the indices for the NPI-B samples that do not pass the Normality test and perform the Wilcoxon test.

Step 1: Perform a preliminary test for Normality on the original sample, with significance level α_1.

Step 2: If the null hypothesis H_0^N of the Normality test is not rejected, then perform the one-sample t-test on the original sample and decide H_0^t with significance level α_2, set $TS^t = 1$ if H_0^t is rejected or set $TS^t = 0$ if H_0^t is not rejected. If H_0^N is rejected, then proceed to perform the Wilcoxon test on the original sample and decide H_0^W with significance level α_2, set $TS^W = 1$ if H_0^W is rejected or $TS^W = 0$ if H_0^W is not rejected.

Step 3: Draw an NPI-B sample N time based on the original sample, with the same size as the original sample.

- For *Cases A* and *B*, perform the Normality test as the preliminary test, followed by either the t-test or the Wilcoxon test according to the Normality test decision. Record the test decision for each iteration, where $TS_i^t = 1$ if H_0^t is rejected for the i-th iteration, or $TS_j^W = 1$ if H_0^W is rejected for the j-th iteration, or record $TS_i^t = 0$ if H_0^t is not rejected or $TS_j^W = 0$ if H_0^W is not rejected, where $i \in N_t^* \subset \{1, 2, \dots, N\}$, and $j \in N_W^* \subset \{1, 2, \dots, N\}$.
- For *Case C*, the t-test is performed on N NPI-B samples if the t-test was applied to the original sample. Record the test decision $T_s = 1$ if H_0^t is rejected for the s-th iteration ($s = 1, 2, \dots, N$), or record $T_s = 0$ if H_0^t is not rejected. The Wilcoxon test is performed on N NPI-B samples if it ends up from the two-stage procedure in the original sample. Record the test decision $W_s = 1$ if H_0^W is rejected for the s-th iteration, or $W_s = 0$ if H_0^W is not rejected.

Step 4: Compute the RP based on the test decisions of the NPI-B samples.

(i) The RP for *Case A*:

If the original sample passed the Normality test and the t-test was applied, then RP is

$$RP_t = \sum_{i \in N_t^*} \mathbb{I}_{\{TS^t = TS_i^t\}} \frac{1}{N}$$

where $\mathbb{I}_{\{TS^t = TS_i^t\}}$ is an indicator function that takes the value 1 if the test decision of the i-th NPI-B sample using the t-test matches the test decision of the original one-sample t-test (TS^t), and 0 otherwise.

If the original sample did not pass the Normality test and the Wilcoxon test was applied, then RP is

$$RP_W = \sum_{j \in N_W^*} \mathbb{I}_{\{TS^W = TS_j^W\}} \frac{1}{N}$$

(ii) The RP for *Case B*:

If the original sample passed the Normality test and t-test was applied, then RP is

$$RP_t = \left(\sum_{i \in N_t^*} \mathbb{I}_{\{TS^t = TS_i^t\}} + \sum_{j \in N_W^*} \mathbb{I}_{\{TS^t = TS_j^W\}} \right) \frac{1}{N}$$

If the original sample did not pass the Normality test and the Wilcoxon test was applied, then RP is

$$RP_W = \left(\sum_{i \in N_t^*} \mathbb{I}_{\{TS^W = TS_i^t\}} + \sum_{j \in N_W^*} \mathbb{I}_{\{TS^W = TS_j^W\}} \right) \frac{1}{N}$$

(iii) The RP for *Case C*:

If the original sample passed the Normality test and the t-test was applied, then RP is

$$RP_t = \sum_{s=1}^{N} \mathbb{I}_{\{TS^t = T_s\}} \frac{1}{N}$$

If the original sample did not pass the Normality test and the Wilcoxon test was applied, then RP is

$$RP_W = \sum_{s=1}^{N} \mathbb{I}_{\{TS^W = W_s\}} \frac{1}{N}$$

Step 5: Perform Steps 3 and 4 in total h times, record the outcomes by RP_{t_k}, and RP_{W_k}, where $k = 1, 2, \ldots, h$.

From Step 5: minimum, mean and maximum of RP values are chosen. The inputs of this algorithm include the original sample, α_1, α_2, the number of runs h, and the number of NPI-B samples per run N. We set $h = 100$, $N = 1000$, and $\alpha_1 = \alpha_2 = 0.05$, the reason for choosing the $N = 1000$ and $h = 100$ is mentioned in Section 1.4.5 of Chapter 1.

Flowcharts for an illustrative example of the reproducibility assessment for the two-stage test of one-sample location test for these cases in Appendix C.1.

4.3 Simulation studies for the reproducibility of the location tests with and without Normality test

Simulation studies are conducted to explore the reproducibility of the two-stage procedure testing. In this procedure, the Normality test is applied in the first stage, if the null hypothesis for the Normality test is not rejected, then the one-sample t-test is used; otherwise, the one-sample Wilcoxon signed-rank test is performed. Normality is assessed using the Shapiro-Wilk (SW) test, which introduced and studied its reproducibility in Chapter 2. The Shapiro-Wilk test was chosen because it is often considered one of the best options for testing Normality because its power and its high RP values indicate good reproducibility. Also, it can be applied to a sample size as small as 5. These simulation studies are conducted by implementing the algorithm described in Section 4.2.

Simulation studies are also conducted to estimate the reproducibility of the one-sample location tests, one-sample t-test and one-sample Wilcoxon signed-rank test, without the preliminary test for Normality. This simulation is conducted by applying the NPI-B-RP Algorithm 1, as presented in Section 1.4.5 of Chapter 1.

The null hypothesis for the first stage (Normality test) is that $H_0^N : X$ the population is Normally distributed, and the alternative hypothesis is that $H_1^N : X$ the population is not Normally distributed. The null hypothesis for the second stage test (one-sample location test) is $H_0^2 : \theta = 0$ and the alternative hypothesis is $H_1^2 : \theta \neq 0$, where θ is the location parameter for population, H_0^2 denotes the null hypothesis for the second stage of the one-sample location test which is either the t-test (H_0^t) hypothesis in which case θ is the mean (μ), or the Wilcoxon test (H_0^W) hypothesis in which case θ is the median (η),

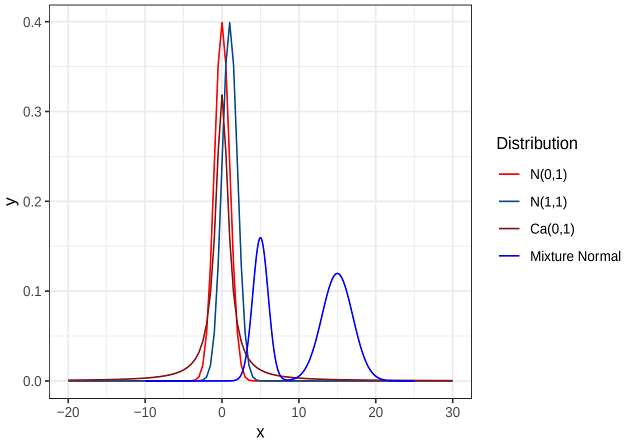


Figure 4.1: The probability density functions (PDFs) of the investigated distributions in the simulation study.

likewise for H_1^2. The hypothesized value for the population mean was chosen equal to 0 because it represents the null hypothesis of no effect.

In these simulation studies, data are simulated under various distributions, each characterized by distinct probability density functions (PDFs). These PDFs are illustrated in Figure 4.1. Specifically, we generate data under the null hypotheses for both stages H_0^N and H_0^2, the standard Normal distribution, denoted by $N(0, 1)$ is chosen for this (Normality and $\mu = 0$). Under the null hypothesis at the first stage and the alternative hypothesis at the second stage H_0^N and H_1^2, we generate data from Normal distributions with a mean of 1 and a standard deviation of 1, denoted by $N(1, 1)$ (Normality and $\mu \neq 0$). Under the alternative hypothesis for the first stage and the null hypothesis at the second stage H_1^N and H_0^2, data are drawn from the standard Cauchy distribution, denoted by $Ca(0, 1)$ (non-Normality and $\eta = 0$). Under the alternative hypothesis at the first and second stage H_1^N and H_1^2, data are drawn from a mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

The number of runs per simulation is $K = 100$, with various sample sizes of 5, 10, 20, and 50, considering different sample sizes allows a precise understanding of how assumptions of Normality affect the reproducibility of the one-sample location tests across various scenarios. The tests are performed for the two-sided hypothesis with a 5% level of significance.

4.3.1 Simulation results for the reproducibility for location tests with preliminary test

In this section, the results of the simulation for NPI-B-RP for one-sample location tests with the preliminary test for Normality are presented. The *Cases A, B* and *C*, introduced in Section 4.2, are considered.

The results are represented visually in plots, where the y-axis represents the min, mean, and max of RP values for the location tests with the preliminary test. In contrast, the x-axis represents the p-values for the location test. The blue colour represents RP values for the two-stage procedure, where the original sample passes the Normality test and performs the t-test. The green represents RP values where the original sample does not pass the Normality test and performs the Wilcoxon test.

The results for *Case A*

This part shows the results of the simulation studies for the full RP value for the two-stage procedure. Generally, from Figures 4.2 - 4.5 that show these results, RP values tend to be low if the p-value for the location test is close to the threshold and when the p-value is far away from the threshold RP values become slightly high.

In scenarios with small sample sizes, it appears that the RP values for the one-sample t-test are slightly higher compared to the RP values for the one-sample Wilcoxon test. However, as the sample size increases, the RP values for the t-test decrease while the RP values for the Wilcoxon test increase. This can be attributed to the ability of the Normality test to detect deviations from Normality, which is weaker in small sample sizes and more robust in larger sample sizes. For small sample sizes, when the Normality test is applied in the first stage, most of the NPI-B samples usually pass the Normality test due to the low power of the Normality test. Consequently, the t-test is performed more than the Wilcoxon test. Thus, the RP for the t-test is higher than the RP for the Wilcoxon test. On the other hand, for larger sample sizes, the Normality test gains more power to detect deviation from Normality in the NPI-B samples, and since NPI-B samples vary in distribution then most of them do not pass the Normality test, leading to the use of the Wilcoxon test instead of the t-test. Thus, the Wilcoxon test tends to have higher RP compared to the t-test for the large sample sizes. Table 4.1 shows an illustrative example, showing that for the small sample size $n = 5$, reveals that N_t, the total number of NPI-B samples passing the Normality test, is greater than N_W, the number of NPI-B samples not passing the Normality test. However, for the larger sample size $n = 50$ from Table 4.2, N_W is greater than N_t. These tables emphasize the effect of sample size on the performance of the Normality test and, subsequently, the choice of location test within the two-stage procedure which ultimately affects their RP. On the other hand, this is related to the strong variability in the RP values, which increase in the t-test as sample size increases, and decrease in the Wilcoxon test as sample size increases. This is because the NPI-B samples are varied in distribution, for the small sample size these samples tend to pass the Normality test and perform the t-test thus there is slight variability in RP for the t-test. As the sample size increases, these samples do not pass the Normality and then apply the Wilcoxon test, thus the Wilcoxon test has low variability in RP values.

Performing the preliminary test for Normality in *Case A* introduces an additional decision-making step in the testing process. This means that there are more opportunities for errors to occur, which can lead to incorrect decisions, which involves incorrectly rejecting a true null hypothesis (Type I error), or failing to reject a false null hypothesis (Type II error), which can adversely affect the RP values. Specifically, when performing the preliminary test for Normality, there is a chance that the test may incorrectly classify a non-Normal sample as Normal or a Normal sample as non-Normal. Such misclassifications can lead to the use of inappropriate statistical tests in the subsequent stage. For example, the t-test might be incorrectly applied to non-Normal data or the Wilcoxon test might be used for data that is Normally distributed. Each of these errors can result in the wrong conclusion about the data, thus impacting the full RP of the two-stage procedure.

When the distribution deviates more from Normality and the sample size is large, all original simulated samples reject the Normality and perform the Wilcoxon test in the second stage. This is reflected in the high reproducibility values observed, as shown in Figures 4.4 and 4.5, particularly for sample sizes of $n = 50$ and when the original samples are drawn from the Cauchy distribution and mixture of Normal distributions.

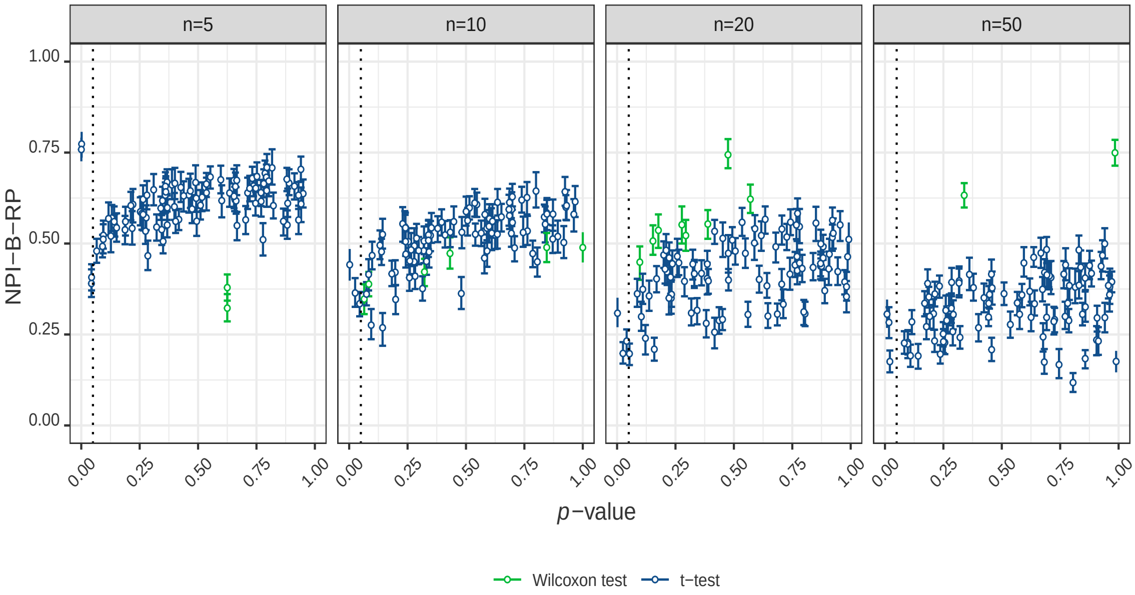


Figure 4.2: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case A*, the original samples are drawn from $N(0,1)$.

Tables 4.1 and 4.2 summarize various outcomes based on the NPI-B samples. In these tables, the notation $N_{p_t \geq \alpha_2}$ represents the number of NPI-B samples that pass the Normality test and have the p-value for the t-test greater than or equal to α_2. Conversely, $N_{p_t < \alpha_2}$ denotes the number of NPI-B samples that pass the Normality test but have the p-value for the t-test less than α_2. Additionally, the tables display the counts for NPI-B samples that did not pass the Normality test. $N_{p_W \geq \alpha_2}$ represents the number of such samples that have the p-value for the Wilcoxon test greater than or equal to α_2, while $N_{p_W < \alpha_2}$ denotes the number of NPI-B samples that did not pass the Normality test with the p-value for the Wilcoxon test less than α_2.

To explain how to evaluate reproducibility in *Case A*, we select four original samples that have p-values close to the threshold and p-values that are very far from the threshold in both areas, which are drawn from $N(0,1)$, considering smallest and largest sample sizes that are studied (5 and 50). This selection is conducted for both the t-test and the Wilcoxon test. The first selected original sample, denoted as $s(N)$, is the original sample that passes the Normality test and shows the smallest p-value for the t-test in the non-rejection area. This choice enables RP to be evaluated in scenarios where the data conform to the assumptions of the t-test and show relatively strong evidence against H_0^2.

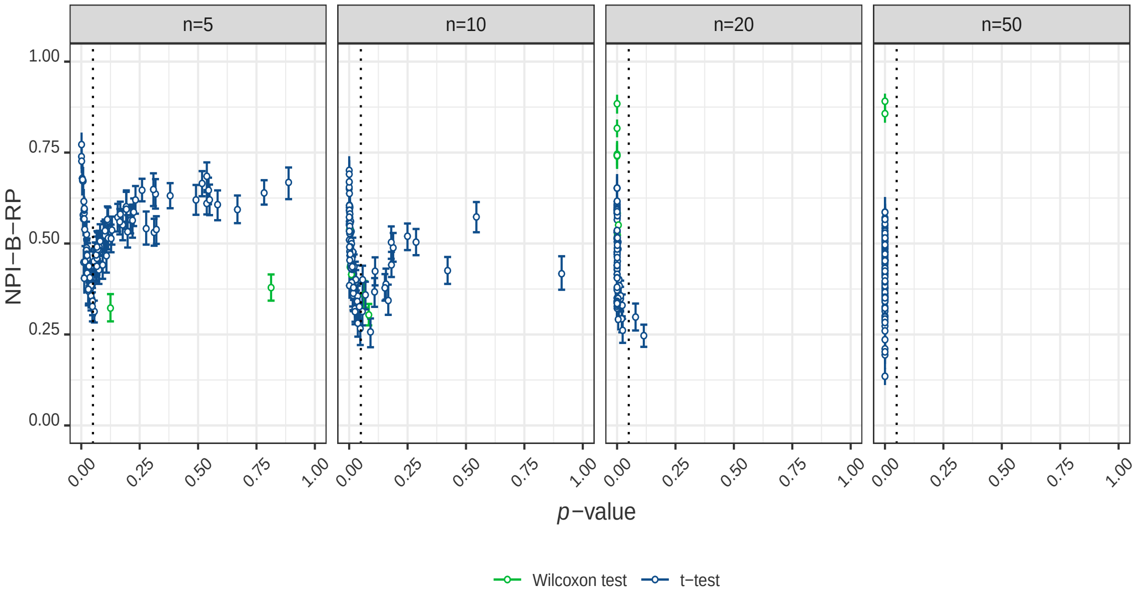


Figure 4.3: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case A*, the original samples are drawn from $N(1,1)$.

The second selected original sample, denoted as $g(N)$, is the original sample that passes the Normality test and has the largest p-value for t-test in the non-rejection area. This choice provides insight into situations where the evidence against H_0^2 is weaker. Moving to samples within the rejection area, the third original sample, denoted as $s(R)$, is the original sample that passes the Normality test but displays the smallest p-value for the t-test. This choice allows RP values to be examined when the evidence against the null hypothesis is relatively stronger in the rejection area. The fourth selected original sample, denoted as $g(R)$, is the original sample that passes the Normality test and has the largest p-value for t-test in the rejection area. This can explore scenarios where the evidence against the null hypothesis is weaker in the rejection region. Similarly, for the Wilcoxon test, the original samples that do not pass the Normality test are selected.

For the $s(N)$ for the t-test, the RP in *Case A* is calculated as the ratio of the total number of NPI-B samples that passed the Normality test and have p-values for the t-test greater than or equal to α_2 to the total number of NPI-B samples. For sample size $n = 5$, $N_{p_t \geq \alpha_2}$ is equal to 490 out of 1000 NPI-B samples. To compute its RP, 490 is divided by the total number of NPI-B samples, $N = 1000$, resulting in an RP of 0.490. For sample

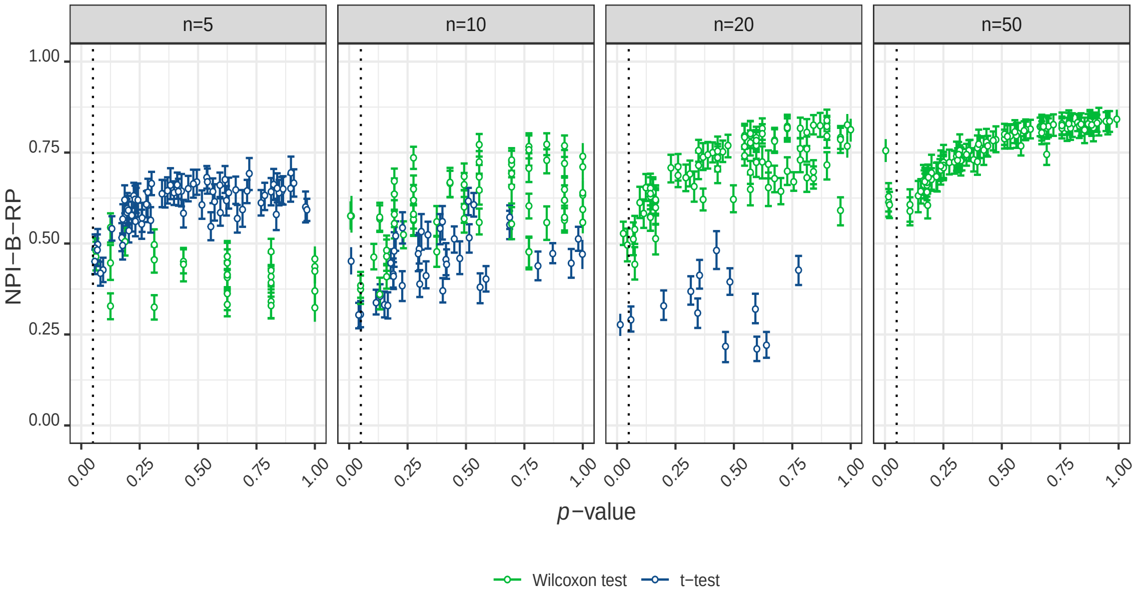


Figure 4.4: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case A*, the original samples are drawn from $Ca(0, 1)$.

size $n = 50$, $N_{p_t \geq \alpha_2}$ is equal to 220, leading to an RP of 0.220. Similarly, RP for this case is calculated for the $s(N)$ for the Wilcoxon test. For sample size $n = 5$, $N_{p_W \geq \alpha_2}$ are equal to 327 NPI-B samples, then RP is 0.327. For sample size $n = 50$, $N_{p_W \geq \alpha_2}$ is equal to 619 NPI-B samples divided by the total number of NPI-B samples giving RP = 0.619. For the $g(N)$ for the t-test, and for sample size $n = 5$, $N_{p_t \geq \alpha_2}$ is equal to 633, then RP is 0.490. For sample size $n = 50$, $N_{p_t \geq \alpha_2}$ is equal to 169, leading to an RP of 0.169. Similarly, RP for this case is calculated for the $g(N)$ for the Wilcoxon test. For sample size $n = 5$, $N_{p_W \geq \alpha_2}$ are equal to 373, then RP is 0.373. For sample size $n = 50$, $N_{p_W \geq \alpha_2}$ is equal to 751 NPI-B samples, then RP = 0.751. For the $s(R)$ for the t-test, and sample size $n = 5$, $N_{p_t < \alpha_2}$ is equal to 754, then RP is 0.754. For sample size $n = 50$, $N_{p_t < \alpha_2}$ is equal to 290, leading to an RP of 0.290. Similarly, RP for this case is calculated for the $g(R)$ for the t-test. For sample size $n = 5$, $N_{p_t < \alpha_2} = 418$, then RP is 0.418. For sample size $n = 50$, $N_{p_t < \alpha_2} = 168$ NPI-B samples, then RP = 0.168.

The results for *Case B*

This case shows the RP values for the same outcome for the location tests, no matter which test is used. Figures 4.6 - 4.9 show the results of simulation for these RP. The RP

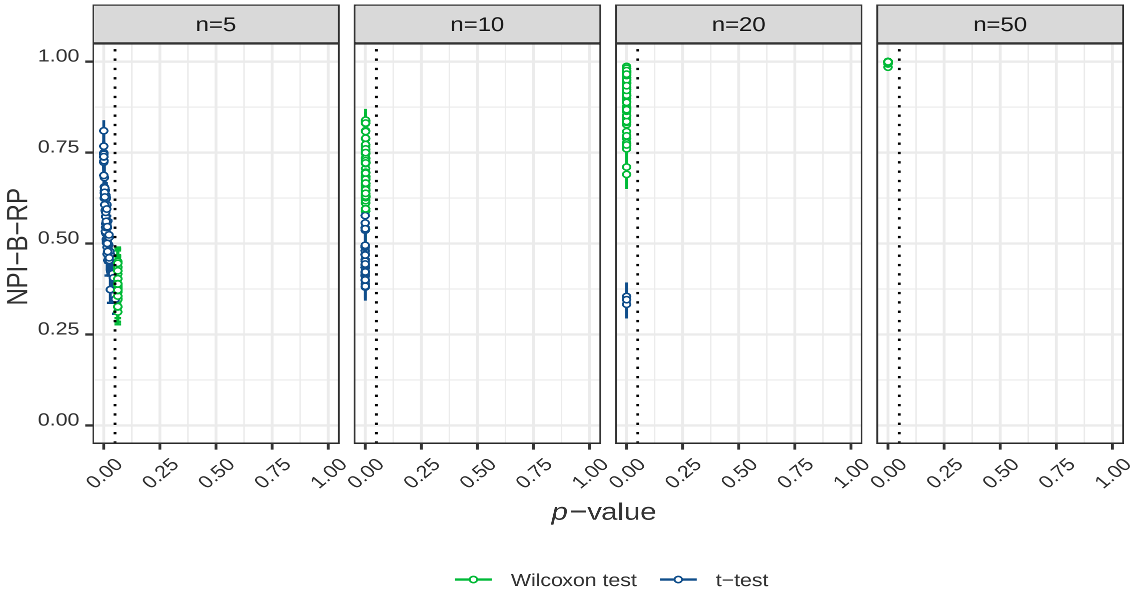


Figure 4.5: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case A*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

values show the general pattern: RP is low when the p-value for location tests is close to the threshold, and RP values are high as the p-value moves away from the threshold. The RP values in the non-rejection area tend to decrease slightly as the sample size increases. Whereas in the rejection area, RP values increase as the sample size increases. This is traced back to the test power which increases with increasing the sample size.

When the distribution is Normal, most of the original samples pass the Normality test and perform t-test, as shown in Figures 4.6 and 4.7. When the distribution is non-Normal, most of the original samples perform the Wilcoxon test, especially for large sample sizes, as shown in Figure 4.8.

Figures 4.6 and 4.8 show RP values under the null hypothesis for location tests. RP values are higher than 50% in the non-rejection area. However, they do not reach close to one when the p-value is close to one, this is because both tests in the two-stage are performed for the two-sided test.

Figure 4.7 shows RP values under the alternative hypothesis for location tests. For sample size $n = 5$, most RP values in the rejection area are less than 50%. Thus, the results

	t-test				Wilicoxon test			
	$s(N)$	$g(N)$	$s(R)$	$g(R)$	$s(N)$	$g(N)$	$s(R)$	$g(R)$
$N_{p_t \geq \alpha_2}$	**490**	**633**	4	337	463	459	NA	NA
$N_{p_t < \alpha_2}$	312	114	**754**	**418**	210	168	NA	NA
N_t	802	747	758	755	673	627	NA	NA
$N_{p_W \geq \alpha_2}$	198	253	242	245	**327**	**373**	NA	NA
$N_{p_W < \alpha_2}$	0	0	0	0	0	0	NA	NA
N_W	198	253	242	245	327	373	NA	NA
RP	0.49	0.633	0.754	0.418	0.327	0.373	NA	NA

Table 4.1: NPI-B sample counts for non-rejection and rejection cases for H_0^t and H_0^W with a preliminary test for Normality. $s(N)$ and $g(N)$ denote the original sample with the smallest and greatest p-values in the non-rejection area, while $s(R)$ and $g(R)$ represent the original sample with the smallest and greatest p-values in the rejection area, respectively. $N_{p_t \geq \alpha_2}$ is the number of NPI-B samples passing the Normality test with a p-value for the t-test greater than or equal to α, $N_{p_t < \alpha_2}$ is the number passing the Normality test with a p-value for the t-test less than α, $N_{p_W \geq \alpha_2}$ is the number with a p-value for the Wilcoxon test greater than or equal to α, and $N_{p_W < \alpha_2}$ is the number with a p-value for the Wilcoxon test less than α, samples from $N(0,1)$ with $n = 5$.

are not very reproducible this is because NPI-B samples that did not pass the Normality test and perform the Wilcoxon test cannot reject the null hypothesis for the Wilcoxon test (this will be explained the reason in Subsection 4.3.2). This leads to a decrease in the number of NPI-B samples that have the same result as the original sample, which rejects the null hypothesis for the t-test, thereby obtaining low RP values. In contrast, RP values in the non-rejection area are higher than 0.6, which indicates reproducibility. As the sample size increases, RP values in the rejection area increase, whereas RP values in the non-rejection area decrease.

To explain how to evaluate reproducibility in *Case B*, we select four original samples for both location tests as presented in *Case A* in Section 4.3.1. Table 4.3 shows various outcomes based on the NPI-B samples for these four original samples that are drawn from $N(0,1)$ with sample size $n = 20$. For the original sample $s(N)$ that passes the Normality test and has the smallest p-value for the t-test in the non-rejection area, the RP value is

	t-test				Wilicoxon test			
	$s(N)$	$g(N)$	$s(R)$	$g(R)$	$s(N)$	$g(N)$	$s(R)$	$g(R)$
$N_{p_t\geq\alpha_2}$	**220**	**169**	172	97	128	76	NA	NA
$N_{p_t<\alpha_2}$	219	33	**290**	**168**	27	22	NA	NA
N_t	439	202	462	265	155	98	NA	NA
$N_{p_W\geq\alpha_2}$	402	658	190	347	**619**	**751**	NA	NA
$N_{p_W<\alpha_2}$	159	140	348	388	226	151	NA	NA
N_W	561	798	538	735	845	902	NA	NA
RP	0.22	0.169	0.29	0.168	0.619	0.751	NA	NA

Table 4.2: NPI-B sample counts for non-rejection and rejection cases for H_0^t and H_0^W with a preliminary test for Normality. $s(N)$ and $g(N)$ denote the original sample with the smallest and greatest p-values in the non-rejection area, while $s(R)$ and $g(R)$ represent the original sample with the smallest and greatest p-values in the rejection area, respectively. $N_{p_t\geq\alpha_2}$ is the number of NPI-B samples passing the Normality test with a p-value for the t-test greater than or equal to α, $N_{p_t<\alpha_2}$ is the number passing the Normality test with a p-value for the t-test less than α, $N_{p_W\geq\alpha_2}$ is the number with a p-value for the Wilcoxon test greater than or equal to α, and $N_{p_W<\alpha_2}$ is the number with a p-value for the Wilcoxon test less than α, samples from $N(0,1)$ with $n = 50$.

computed as all NPI-B samples that pass the Normality test and have the p-values for the t-test greater than or equal to α_2 which is $N_{p_t\geq\alpha_2} = 180$ plus all NPI-B samples that do not pass the Normality but have the p-value for Wilcoxon test greater than or equal to α_2, which are $N_{p_W\geq\alpha_2} = 384$, divide by the total number of NPI-B samples ($N = 1000$), thus $RP = 0.564$. Similarly, for the original sample $s(N)$ that did not pass the Normality test and has the smallest p-value for the Wilcoxon test in the non-rejection area, the RP value is calculated as the ratio of all NPI-B samples that do not pass the Normality test and have the p-value for Wilcoxon test greater than or equal to α_2 which are $N_{p_W\geq\alpha_2} = 439$ plus all NPI-B samples that pass the Normality test and have the p-values for the t-test greater than or equal to α_2 which are $N_{p_t\geq\alpha_2} = 150$ to the total number of NPI-B samples. Likewise, the reproducibility is calculated for other samples.

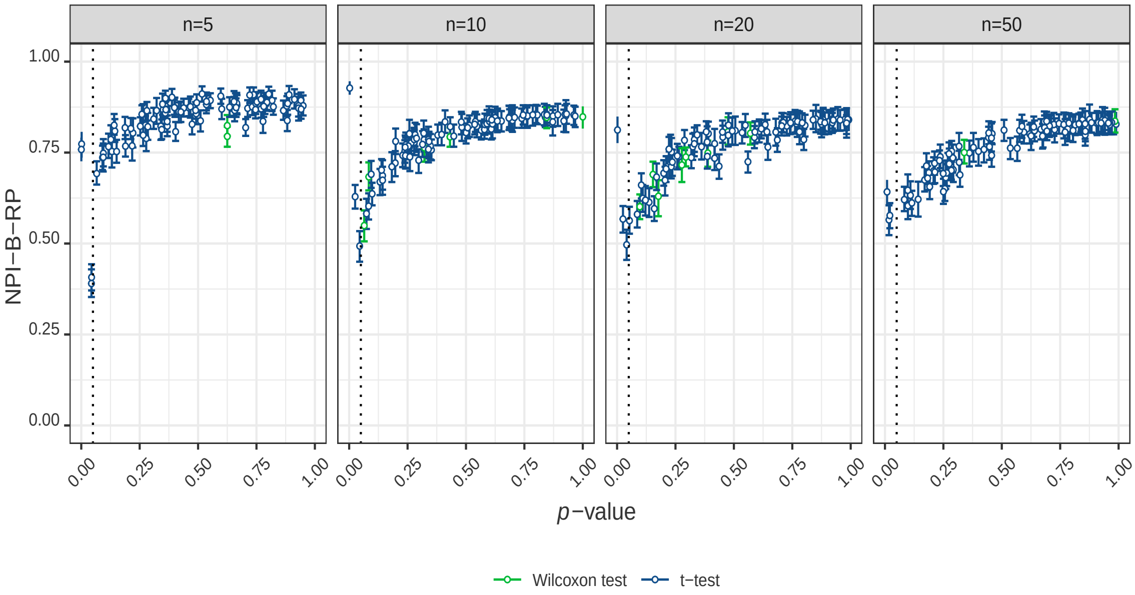


Figure 4.6: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case B*, the original samples are drawn from $N(0, 1)$.

The results for *Case C*

This part shows the results of simulations for the reproducibility of the location test conclusion, where for the NPI-B samples the same location test is applied as for the original sample without further a preliminary test. Generally, from Figures 4.10 - 4.13 that show these results for different distributions and sample sizes, RP values tend to be low if the p-value for the location test is close to the threshold, while when the p-value is far away from the threshold RP values become high. The results are similar to those observed in *Case B*, except that the RP for original samples performing the Wilcoxon test with a sample size of 5 always equals one. Additionally, for a sample size of 5 under H_1^2, when original samples are drawn from $N(1, 1)$, most of these original samples have RP values for the t-test greater than 0.5, as shown in Figure 4.11. This is because H_0 for the Wilcoxon test is not rejected.

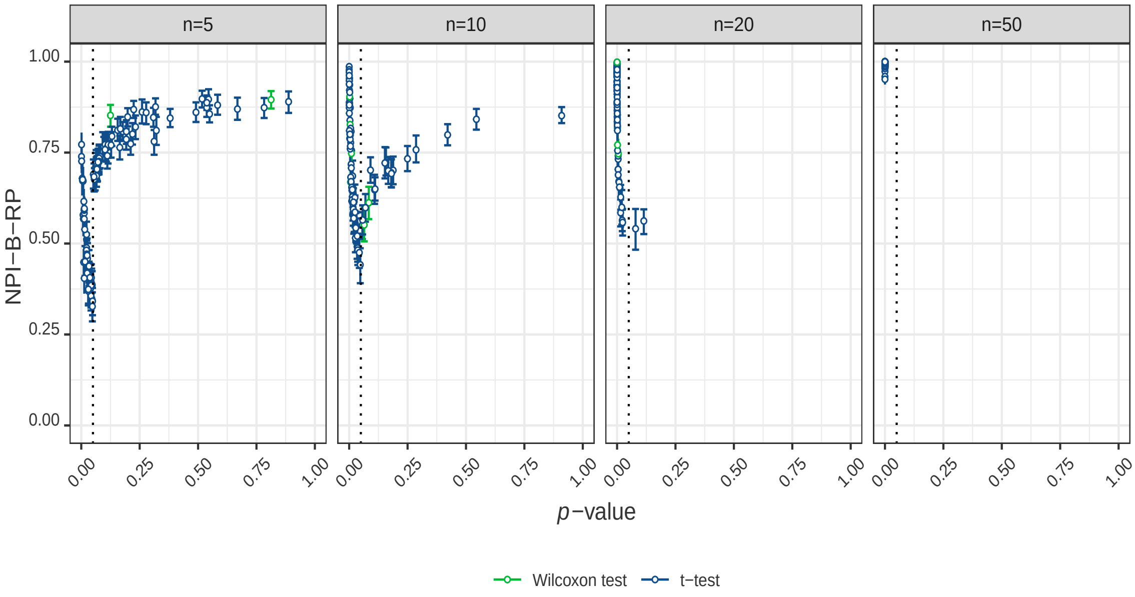


Figure 4.7: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case B*, the original samples are drawn from $N(1,1)$.

The relationship between RP and the estimated power for the two-stage procedure

Table 4.4 illustrates the relationship between the overall mean of RP values in the rejection area and the estimated power for the two-stage procedure. The RP values are for samples located in the rejection area out of 100 original samples drawn from the $N(1,1)$ distribution, under the alternative hypothesis. In the two-stage testing procedure, the Shapiro-Wilk (SW) test for Normality is conducted, followed by either the t-test or the Wilcoxon test based on the outcome of the SW test. This procedure is applied to both the original sample and N NPI-B samples. RP represent the proportion of NPI-B samples where the outcome of either test (t-test or the Wilcoxon test) matches that of the original sample's location test, divided by the total number of NPI-B samples.

A Monte Carlo simulation of $10,000$ datasets is performed to estimate the power of the two-stage testing procedure, where the SW test is conducted in the first stage and then proceeds with either the t-test or the Wilcoxon test based on the outcome of the SW test. The power will be the proportion of datasets for which either test successfully rejects the null hypothesis H_0^2 to the total number of the datasets.

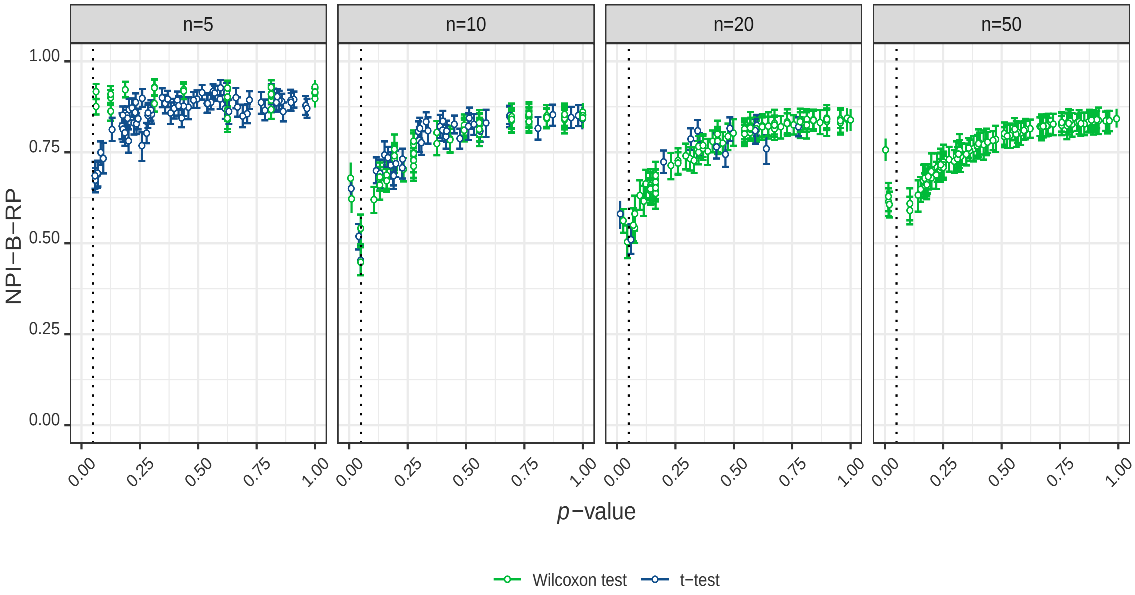


Figure 4.8: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case B*, the original samples are drawn from $Ca(0, 1)$.

Table 4.4 shows that RP increases as the estimated power increases. This indicates that as tests become better able to detect true effects, the reproduction of these results in future studies also increases. Furthermore, as the sample size increases, both the estimated power and RP increase. This is because larger samples offer more data, enhancing the ability to detect true effects and consequently boosting the reproducibility of the results.

4.3.2 Simulation results for the reproducibility of location tests without preliminary test

This subsection presents the outcomes obtained from the simulation studies focused on estimating the RP using the NPI-B-RP method for the one-sample location tests without the preliminary test of Normality. These RP values were studied to explore the effect of the Normality test on RP of the location tests by comparing RP of these results with RP values for location tests with the preliminary test of Normality which were presented in the previous subsection.

The RP values of the one-sample t-test without the preliminary test for Normality are illustrated in Figures 4.14 - 4.17. The RP values for the t-test without the preliminary

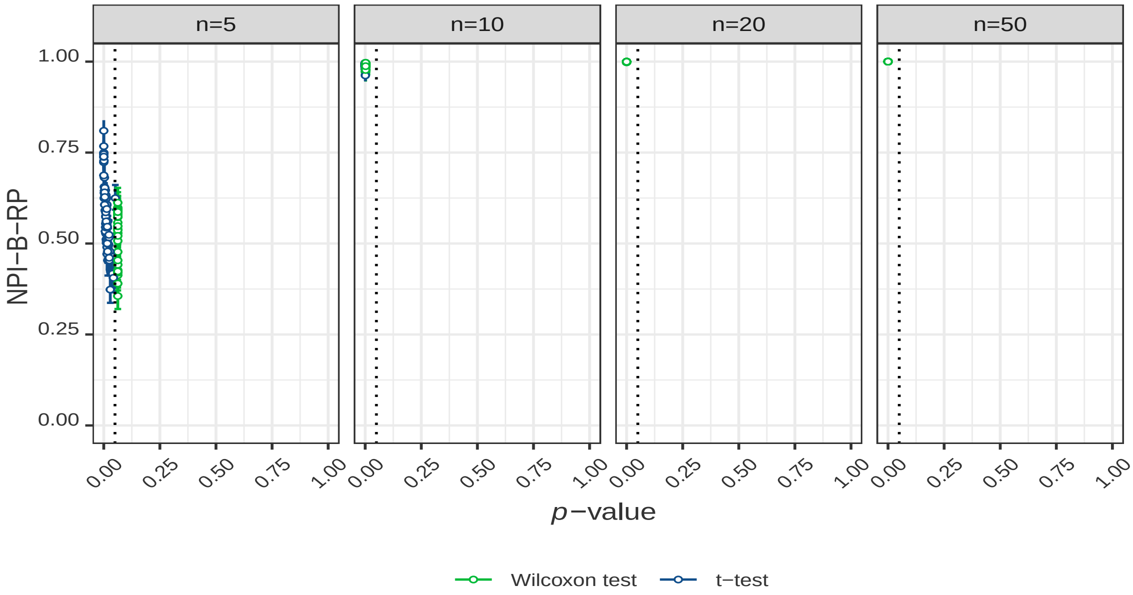


Figure 4.9: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case B*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

test follow the general pattern for RP: when the p-value is close to the threshold the RP values tend to be small, whereas if the p-value is far away from the threshold the RP increases.

Figures 4.18 - 4.21 show the RP values for the Wilcoxon test without the preliminary test for Normality. Generally, it can be noticed that the RP values for the Wilcoxon test without performing the preliminary test for Normality also follow the general pattern for RP. However, this pattern of RP is not true for sample size 5 where RP values for all original samples are equal to one. This is because for the two-sided Wilcoxon test with sample size $n = 5$ and $\alpha = 0.05$ the null hypothesis is never rejected because the smallest possible two-sided p-value with $n = 5$ is 0.0625, which is greater than the significance level $\alpha = 0.05$. Specifically, for a sample size of 5, there are $2^n = 2^5 = 32$ possible rankings, the probability of observing the most extreme case under the null hypothesis (all 5 observations being either greater than or less than the hypothesized median) is $\frac{1}{32}$. For a two-sided test, the probability will be doubled to account for both sides, thereby the smallest possible two-sided p-value is $2 \cdot \frac{1}{32} = 0.0625$. Consequently, the null hypothesis cannot be rejected even when a true difference exists [43]. For this reason, the RP values

	t-test				Wilicoxon test			
	$s(N)$	$g(N)$	$s(R)$	$g(R)$	$s(N)$	$g(N)$	$s(R)$	$g(R)$
$N_{p_t \geq \alpha_2}$	**180**	**513**	86	232	**150**	**183**	NA	NA
$N_{p_t < \alpha_2}$	151	66	**303**	**228**	103	40	NA	NA
N_t	331	579	389	460	253	223	NA	NA
$N_{p_W \geq \alpha_2}$	**384**	**337**	98	298	**439**	**615**	NA	NA
$N_{p_W < \alpha_2}$	285	84	**513**	**242**	308	162	NA	NA
N_W	669	421	611	540	747	777	NA	NA
RP	0.564	0.85	0.816	0.470	0.589	0.798	NA	NA

Table 4.3: NPI-B sample counts for non-rejection and rejection cases for H_0^t and H_0^W with a preliminary test for Normality. $s(N)$ and $g(N)$ denote the original sample with the smallest and greatest p-values in the non-rejection area, while $s(R)$ and $g(R)$ represent the original sample with the smallest and greatest p-values in the rejection area, respectively. $N_{p_t \geq \alpha_2}$ is the number of NPI-B samples passing the Normality test with a p-value for the t-test greater than or equal to α, $N_{p_t < \alpha_2}$ is the number passing the Normality test with a p-value for the t-test less than α, $N_{p_W \geq \alpha_2}$ is the number with a p-value for the Wilcoxon test greater than or equal to α, and $N_{p_W < \alpha_2}$ is the number with a p-value for the Wilcoxon test less than α, samples from $N(0, 1)$ with $n = 20$.

for the Wilcoxon test for $n = 5$ are always equal to one which means there are $N = 1000$ NPI-B samples of $N = 1000$ NPI-B samples that have p-values greater than α.

Generally, the results of simulation for RP values for the one-sample location tests without the preliminary test for Normality show that the RP tends to be low when the p-value for the location test is close to α. When the p-value is close to α, it suggests weak evidence against or for H_0. This implies that the p-value for the test in future experiments is not equally likely to be similar to the p-value for the original test, resulting in low RP values. Conversely, when the p-value is far away from α, the RP tends to be high. This is because when the p-value is far from the threshold, indicating strong evidence against or for H_0, it means the p-value for the test in future experiments is equally likely to be similar to the p-value for the original test, thus leading to high RP values. Moreover, the RP values do not reach one when the p-values are close to one this is because the

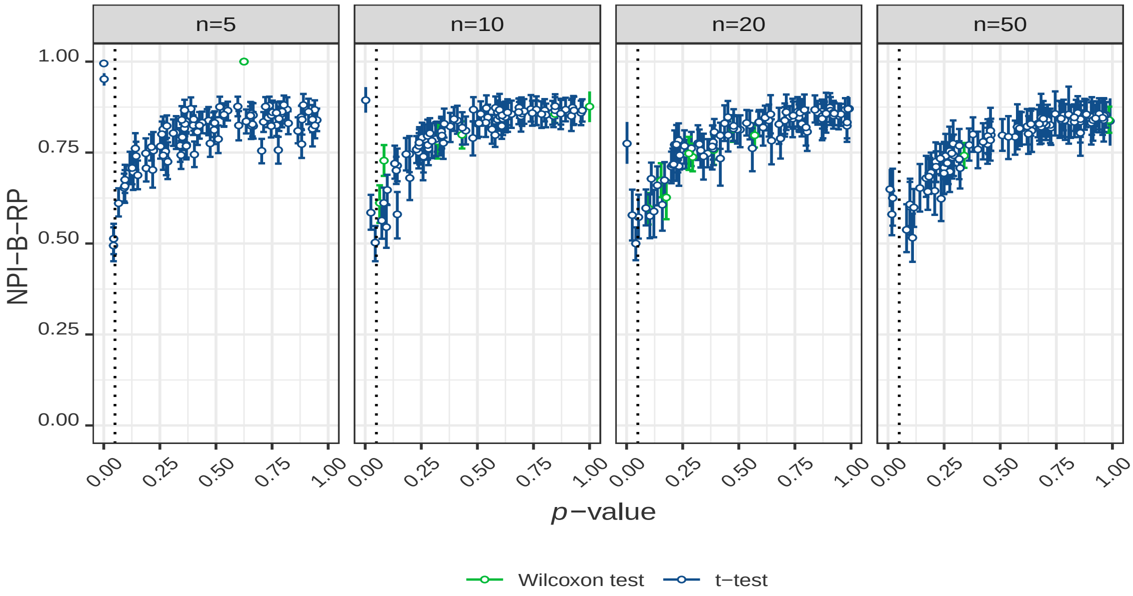


Figure 4.10: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case C*, the original samples are drawn from $N(0, 1)$.

Sample size	RP	Power
$n = 5$	0.499	0.380
$n = 10$	0.769	0.807
$n = 20$	0.934	0.990
$n = 50$	0.997	0.999

Table 4.4: The relationship between the overall mean of RP values in the rejection area and the estimated power for the two-stage procedure, samples from $N(1, 1)$.

one-sample location tests are performed for the two-sided test.

When comparing the reproducibility for the t-test and the Wilcoxon test, there is no substantial difference observed in RP values when conducting them without the preliminary test for Normality for sample sizes of 10, 20 and 50. This is illustrated in Figures 4.22, 4.23 and 4.24. However, when data are sampled from non-Normal distributions, slight variability in the RP of the t-test is observed. This is because the parametric t-test is particularly sensitive to violations of Normality assumptions, resulting in greater variability in its RP values. In contrast, the nonparametric Wilcoxon test is more robust to these violations, leading to more consistent RP values.

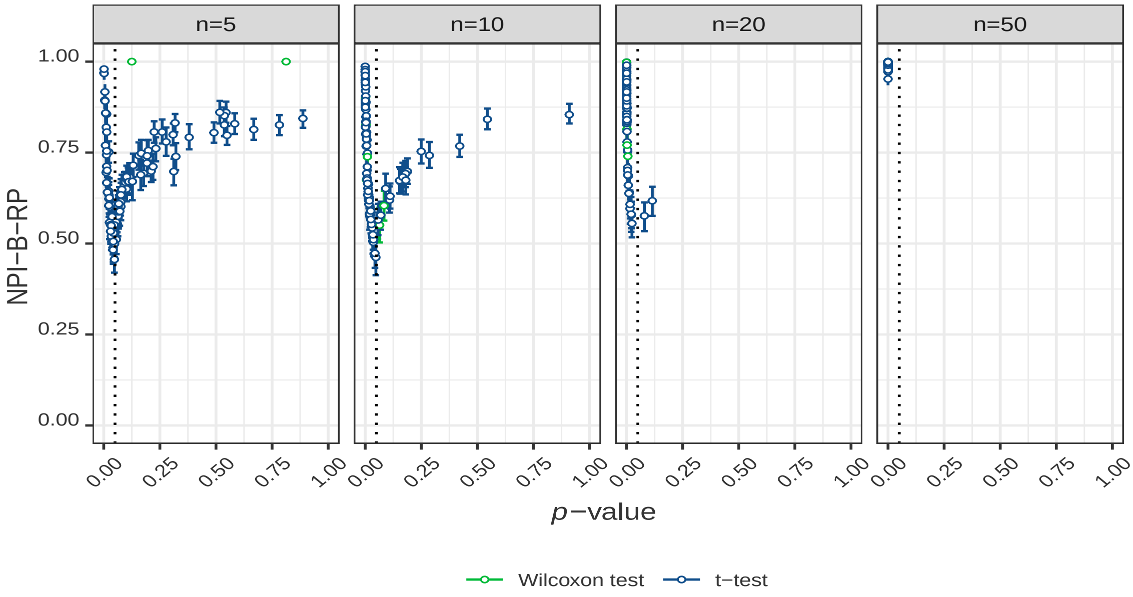


Figure 4.11: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case C*, the original samples are drawn from $N(1, 1)$.

When data are simulated under the null hypothesis of location tests the most original samples are located in the non-rejection area as shown in Figures 4.14, 4.16, 4.18 and 4.20. Whereas when data are simulated under the alternative hypothesis of location tests, most original samples are located in the rejection area, increasing the number as the sample size increases as shown in Figures 4.15 and 4.19.

The relationship between RP and the estimated power for the one-sample location tests without preliminary test

The relationship between the overall mean of RP values in the rejection area and the estimated power for one-sample location tests without preliminary tests is examined. Table 4.5 shows this relationship under the alternative hypothesis when samples are drawn from $N(1, 1)$. The estimated power is calculated via Monte Carlo simulation for the 10,000 dataset. For each simulated dataset, we perform the location test. Then determine how often the test correctly rejects H_0. The proportion of times H_0 is rejected out of the total number of simulations provides an estimate of the power of the location test.

It is evident from the table that as the power increases, the RP also increases. Moreover, with larger sample sizes, both RP in the rejection area and power tend to

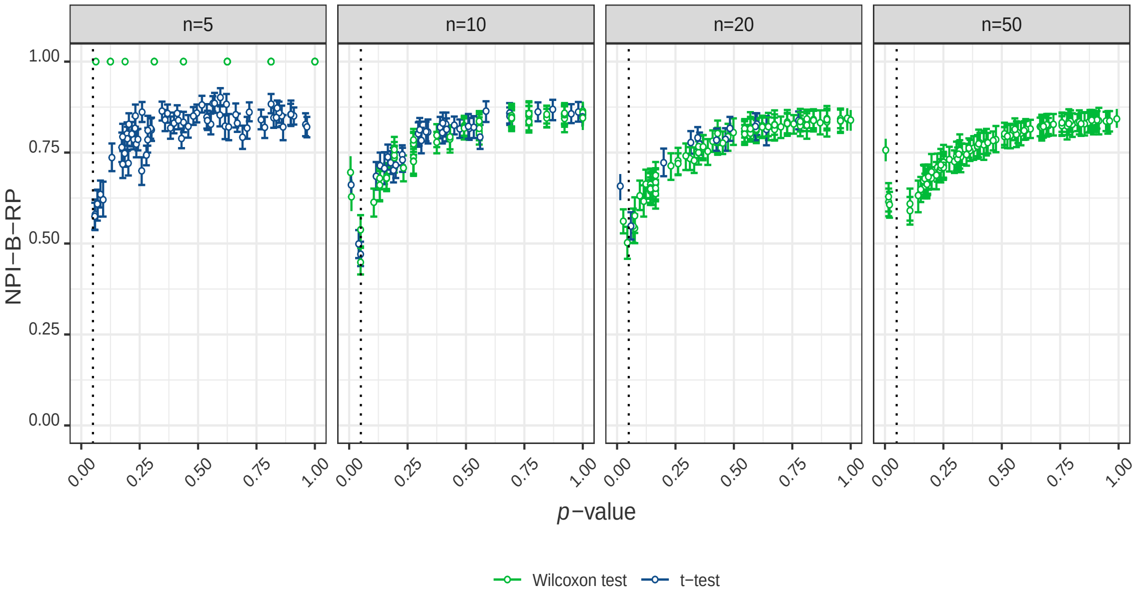


Figure 4.12: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case C*, the original samples are drawn from $Ca(0, 1)$.

increase. For sample size $n = 50$, both tests have the same power and RP, whereas, for the sample sizes $n = 5, 10, 20$, the t-test has slightly better power and RP than the Wilcoxon test. This is because the distribution is Normal.

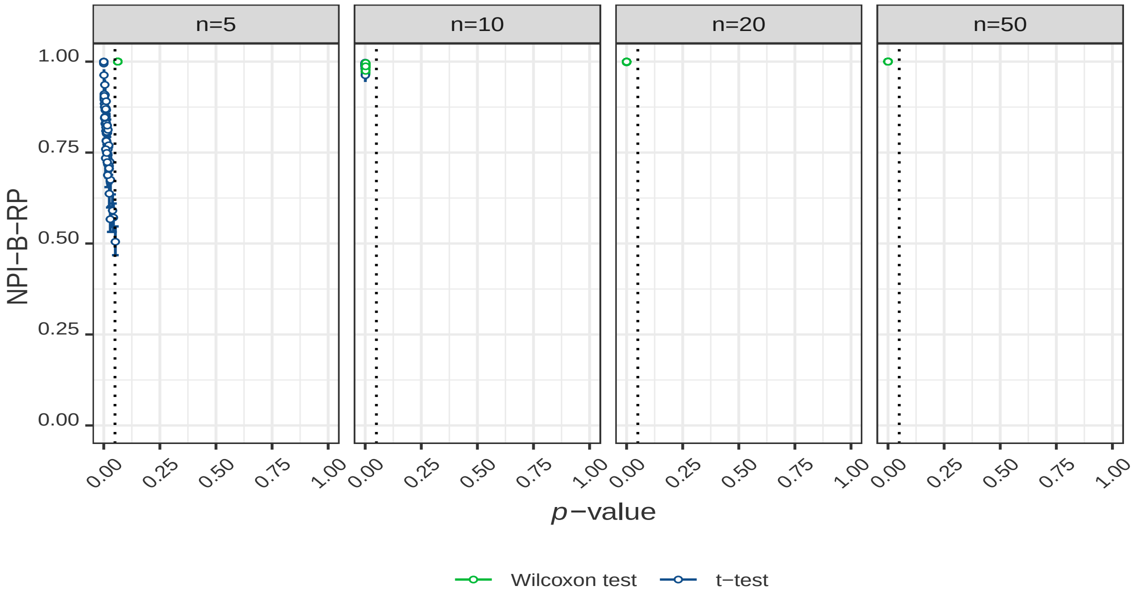


Figure 4.13: The min, mean, and max of RP values against the *p*-values for the two-stage procedure for *Case C*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

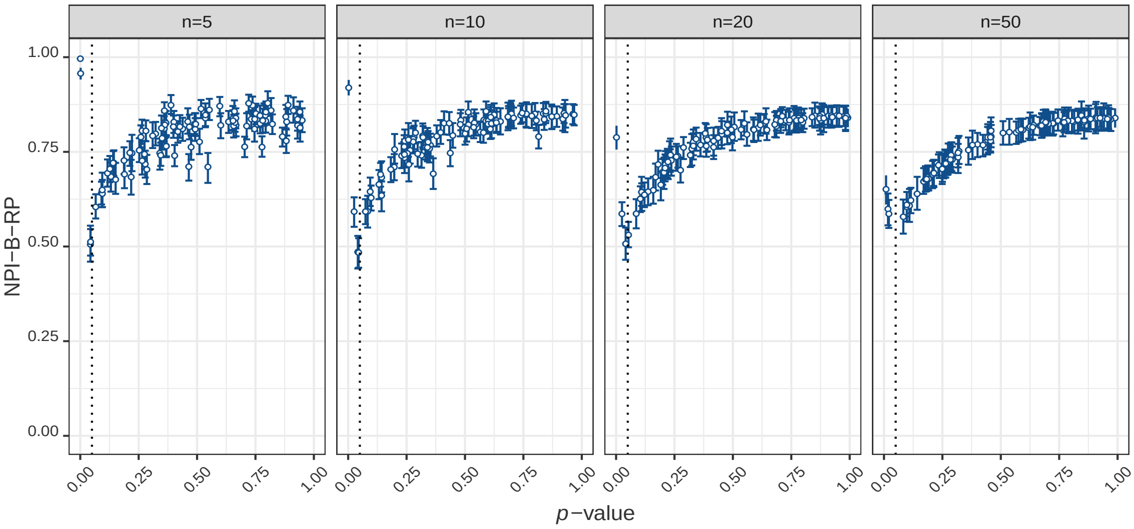


Figure 4.14: The min, mean and max of RP values against the *p*-values for the one-sample *t*-test, with original samples from $N(0, 1)$.

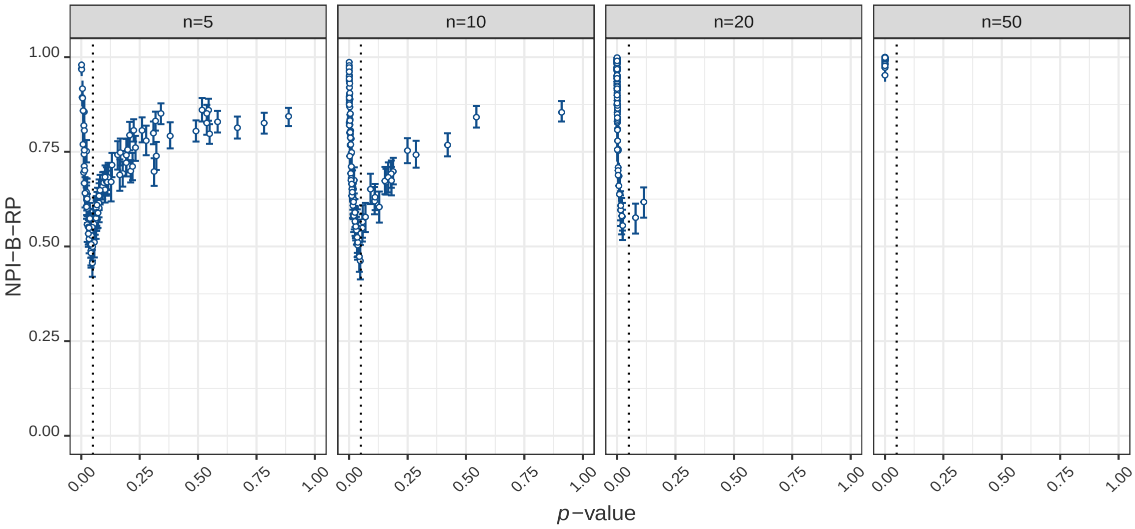


Figure 4.15: The min, mean and max of RP values against the p-values for the one-sample t-test, with original samples from $N(1,1)$.

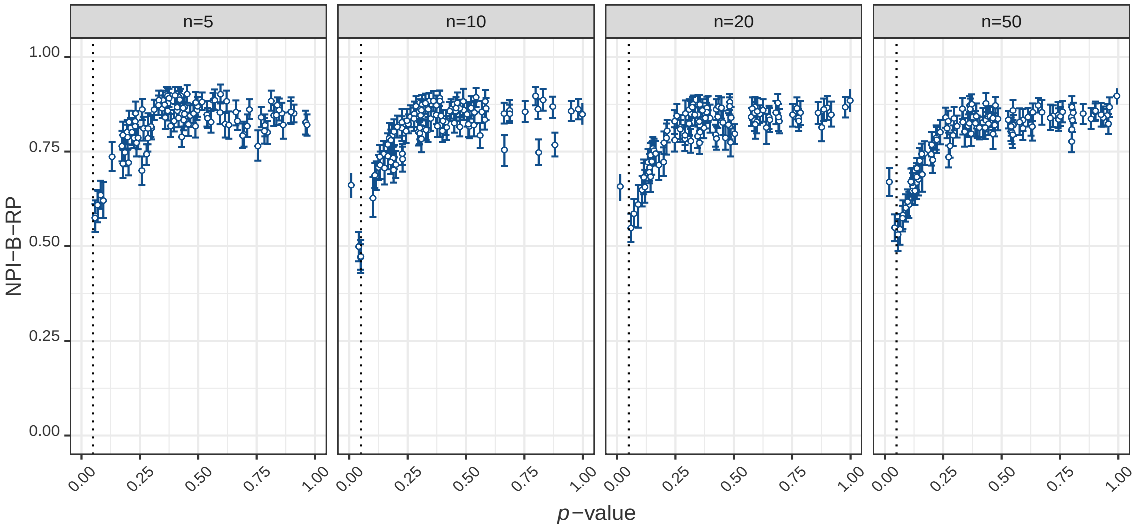


Figure 4.16: The min, mean and max of RP values against the p-values for the one-sample t-test, with original samples from $Ca(0,1)$.

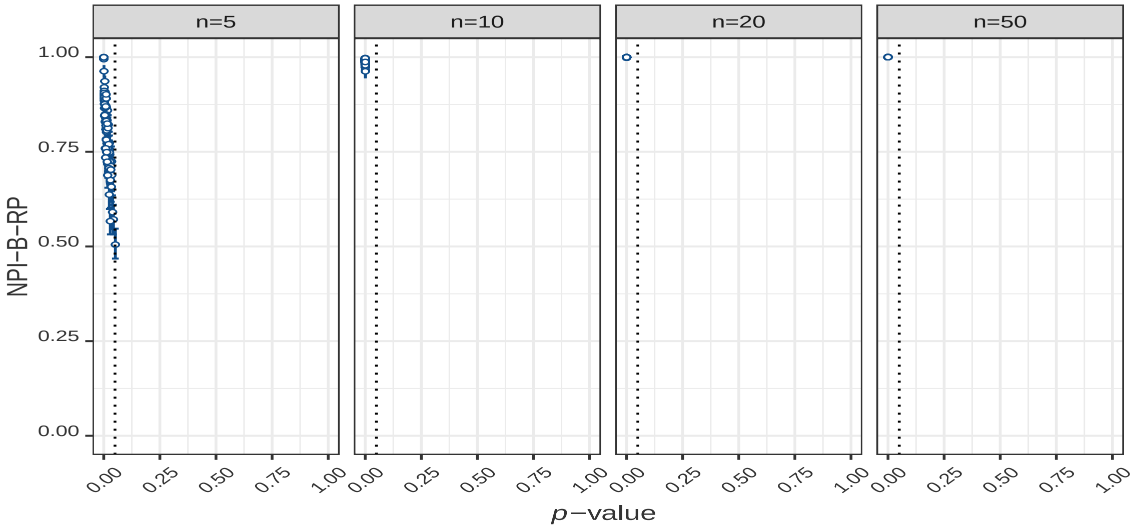


Figure 4.17: The min, mean and max of RP values against the p-values for the one-sample t-test, with original samples from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

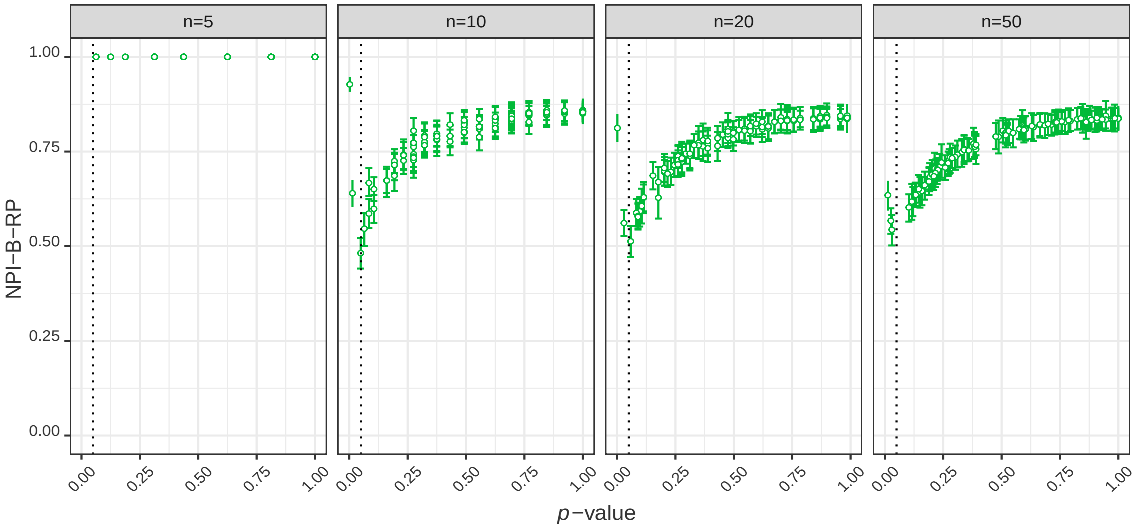


Figure 4.18: The min, mean and max of RP values against the p-values for the one-sample Wilcoxon test, with original samples from $N(0, 1)$.

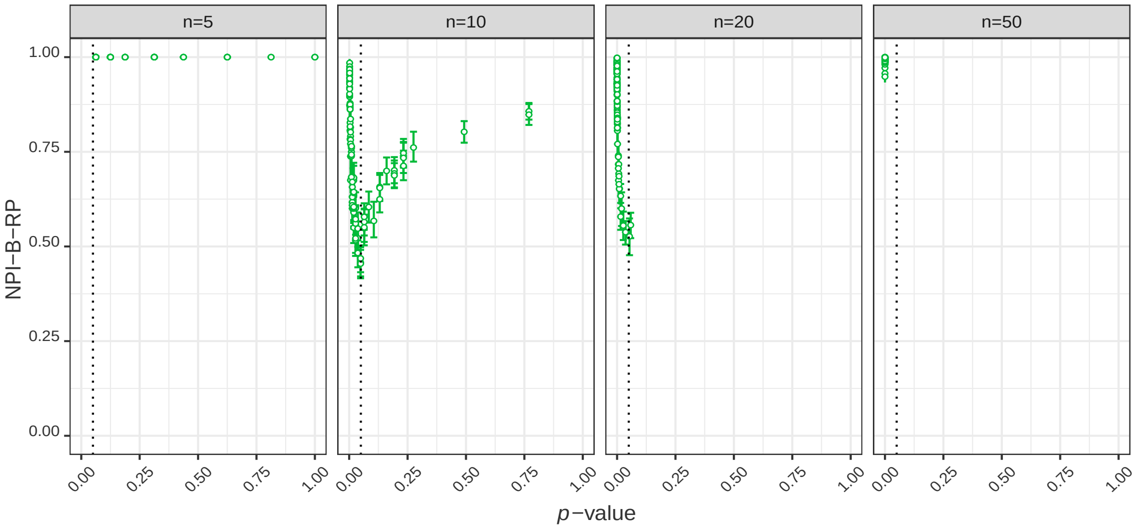


Figure 4.19: The min, mean and max of RP values against the p-values for the one-sample Wilcoxon test, with original samples from $N(1, 1)$.

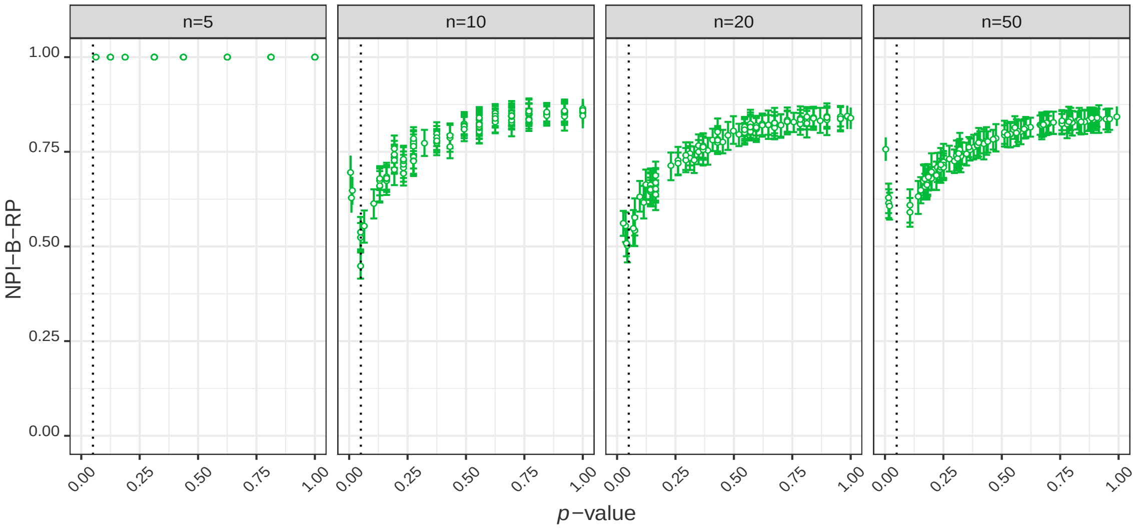


Figure 4.20: The min, mean and max of RP values against the p-values for the one-sample Wilcoxon test, with original samples from $Ca(0, 1)$.

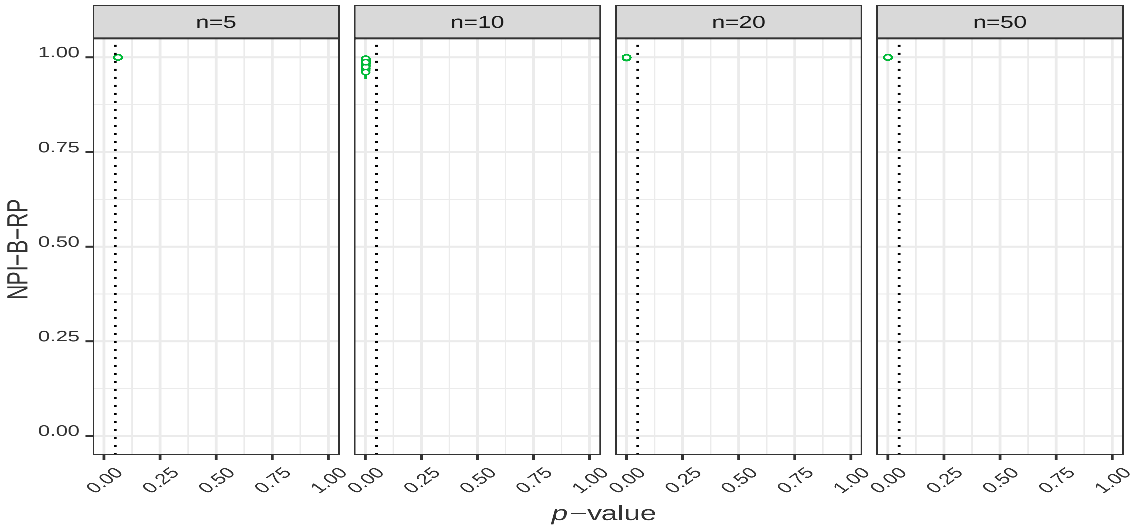


Figure 4.21: The min, mean and max of RP values against the p-values for the one-sample Wilcoxon test, with original samples from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

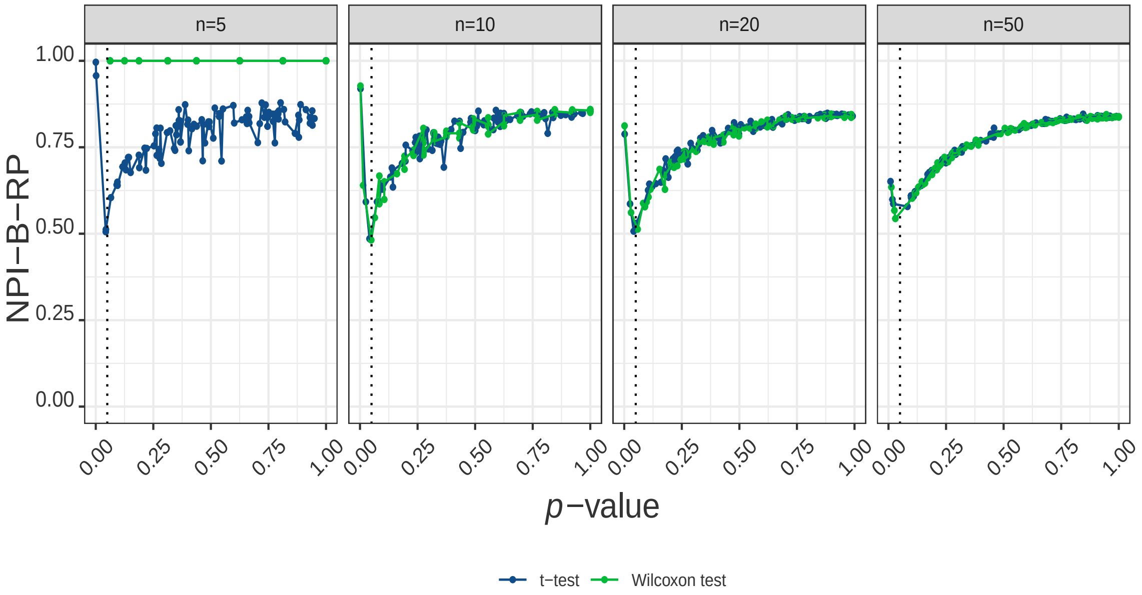


Figure 4.22: The mean of RP values against the p-values for the location tests, with original samples from $N(0, 1)$.

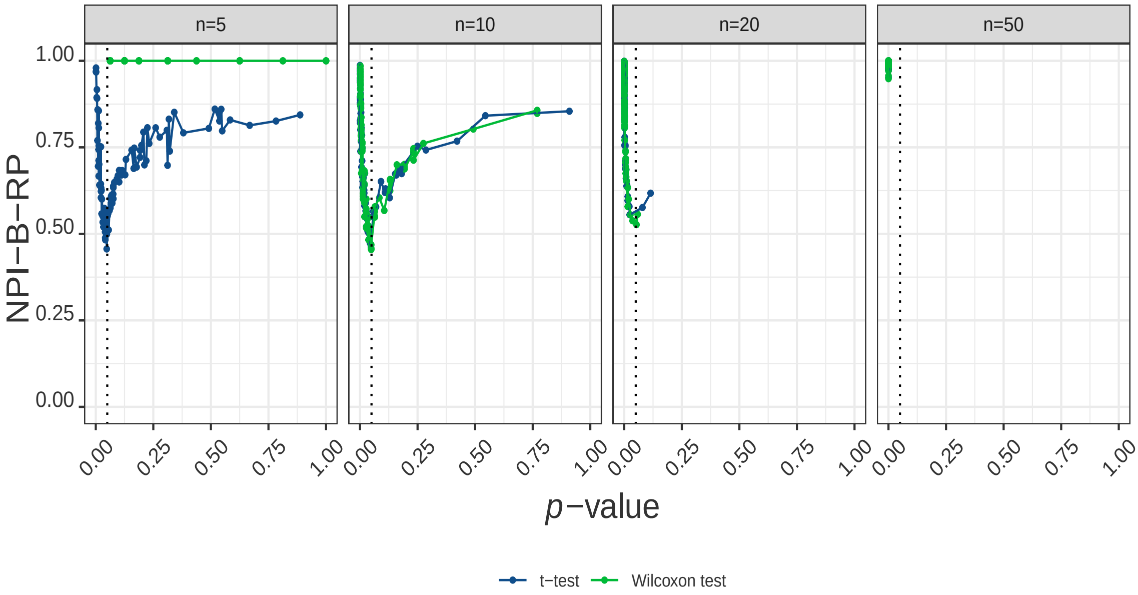


Figure 4.23: The mean of RP values against the p-values for the location tests, with original samples from $N(1, 1)$.

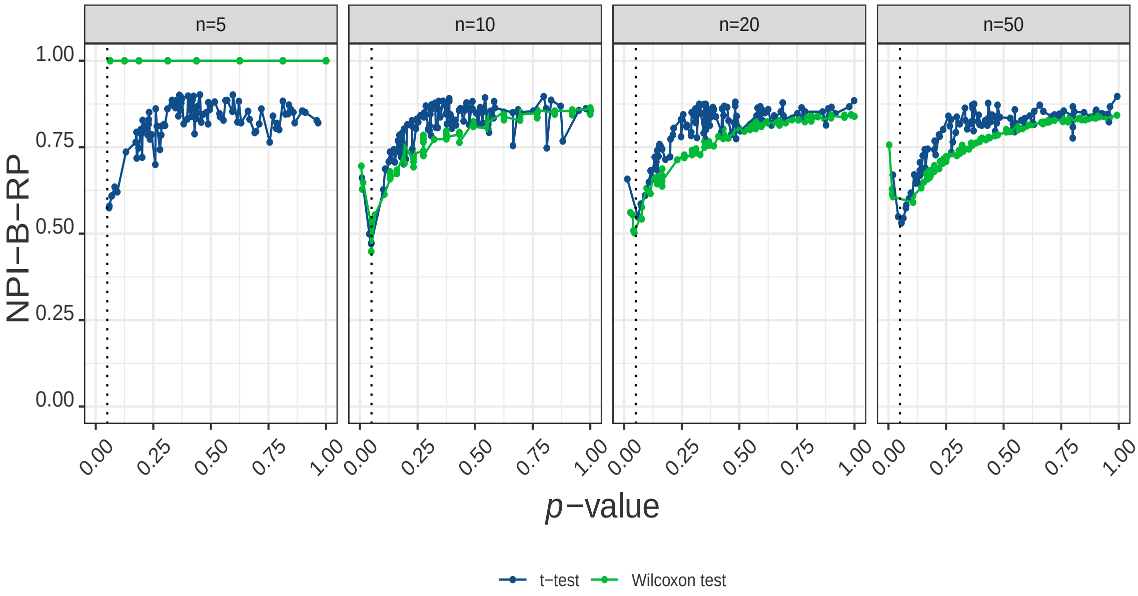


Figure 4.24: The mean of RP values against the p-values for the location tests, with original samples from $Ca(0, 1)$.

Sample size	t-test		Wilicoxon test	
	RP	Power	RP	Power
$n = 5$	0.675	0.401	0.000	0.000
$n = 10$	0.756	0.806	0.750	0.782
$n = 20$	0.886	0.988	0.876	0.984
$n = 50$	0.994	1.000	0.994	1.000

Table 4.5: The relationship between the overall mean of RP values in the rejection area and the power for the location tests without preliminary test, data are sampled from $N(1, 1)$.

4.4 The impact of the preliminary test on reproducibility of location tests

This section assesses the influence of the preliminary test for the Normality on the reproducibility of the one-sample location tests. Does conducting the preliminary test of Normality improve and increase RP values for the location test or not? This is achieved by comparing the RP for the location test with the Normality test with the RP for the location test without the Normality test.

This comparison is done as follows: For *Case A*, the overall mean of RP values for the t-test for this case is compared to the product of the overall mean of the individual RP for the t-test and RP for the Normality test. Similarly, RP for the Wilcoxon test is compared. For *Case B*, the overall mean of RP values for the t-test is compared to the overall mean of the corresponding RP values for the t-test without the preliminary test for Normality. Similarly, the overall mean of RP values for the Wilcoxon test is compared to the overall mean of the corresponding RP values for the Wilcoxon test without the preliminary test for Normality. For *Case C*, the overall mean of RP values for the t-test, in this case, is compared to the overall mean of the RP values for all original samples that perform the t-test without the preliminary test for Normality. similarly, the examination is carried out for the Wilcoxon test.

These comparisons are displayed visually, where the light blue circle which represents the RP values of the t-test with the preliminary test is compared to the dark blue circle indicating the RP values of the t-test without the preliminary test. Similarly, we compare the light green square representing the RP values of the Wilcoxon test with the preliminary test to the dark green square representing the RP values of the Wilcoxon test without the preliminary test. The preliminary test is shortened to the pre-test in the plots.

4.4.1 The impact of the Normality test on RP for location test for *Case A*

Here, the results of comparing the overall mean of RP for *Case A* with the product of the overall mean of the individual RP of location tests and the preliminary test are displayed,

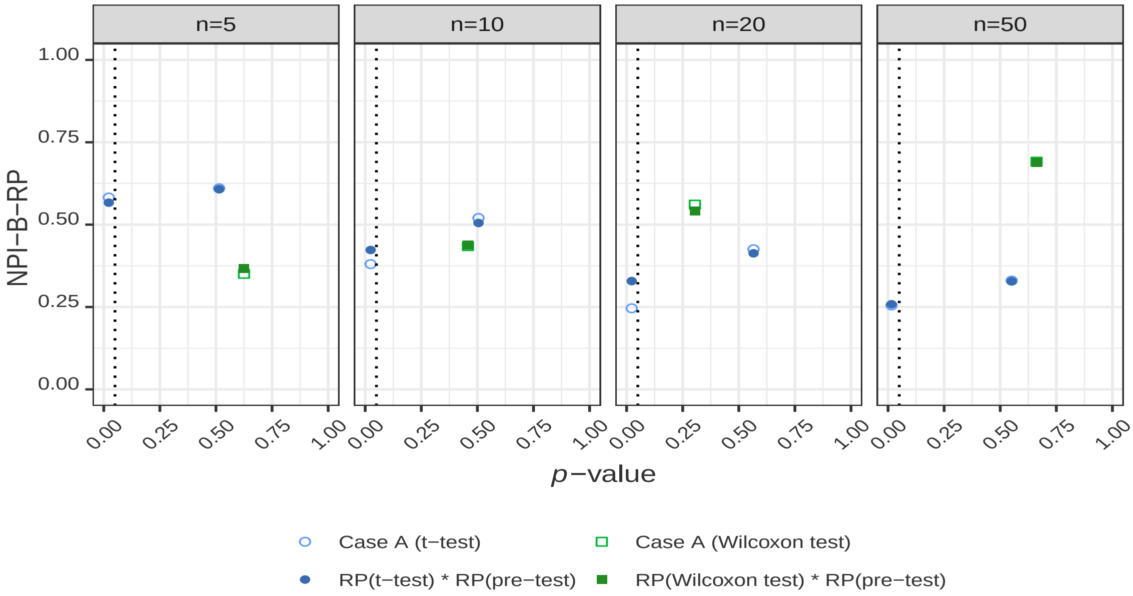


Figure 4.25: Comparing RP values for location tests with and without preliminary test (*Case A*), plotted against their corresponding mean p-values. When the original samples are drawn from $N(0, 1)$.

illustrating the effect of applying the preliminary Normality test on the RP values for the location tests. Figures 4.25 - 4.28 illustrate this comparison for the investigated distributions and sample sizes. The preliminary test of Normality appears to have a very small impact on the RP of the one-sample locations. There is no consistent and clear pattern to this impact. This suggests that the reproducibility of the location test, conditional on the Normality test outcome, is not substantially better than that of the location test alone.

4.4.2 The impact of the preliminary test on RP for location test for *Case B*

In this subsection, we present the results of comparing RP for *Case B*, which explore the potential impact of the preliminary Normality test on the reproducibility of the same outcome for the location tests. This comparison is made with RP values for a location test conducted without the preliminary test. Figures 4.29 - 4.32 show this comparison. The impact of the preliminary test of Normality on RP of the one-sample locations is very

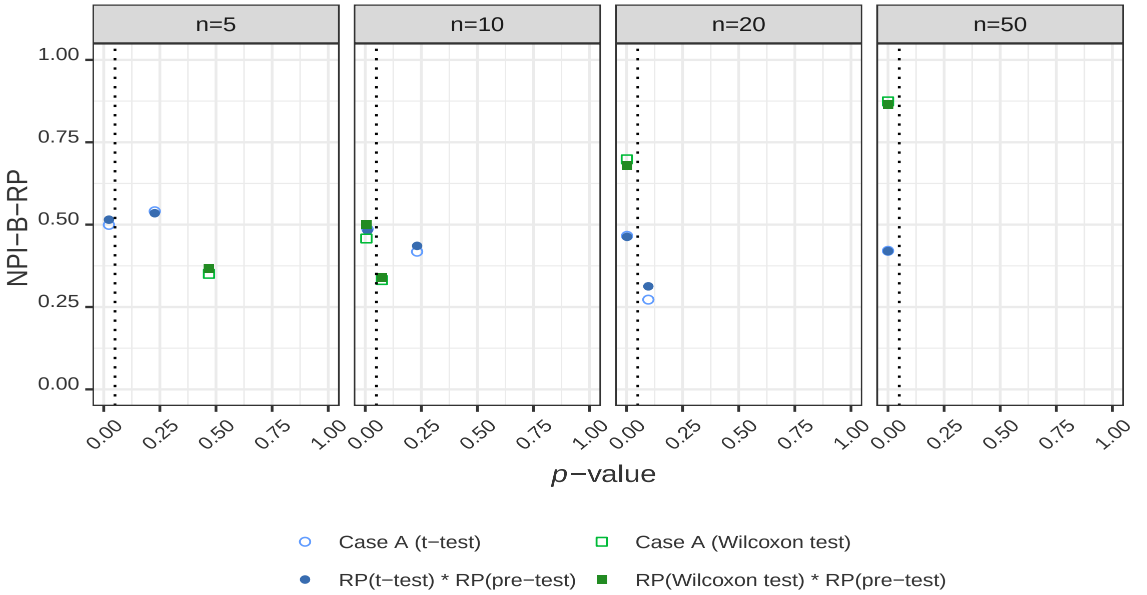


Figure 4.26: Comparing RP values for location tests with and without preliminary test (*Case A*), plotted against their corresponding mean p-values. When the original samples are drawn from $N(1, 1)$.

small, Applying the Normality test does not substantially increase RP of the location test or decrease RP.

The sample size of $n = 5$ is the most affected by applying the Normality test. Where RP for the t-test with preliminary test is slightly higher than RP for the t-test without preliminary test in the non-rejection area. Whereas the opposite happens in the Wilcoxon test, RP for the Wilcoxon test without the preliminary test is higher than RP for the Wilcoxon test with the preliminary test. This is because RP for *Case B* depends on NPI-B samples that have the same outcome as original samples whether perform t-test or Wilcoxon test. NPI-B samples, which initially failed to pass the Normality test and subsequently use the Wilcoxon test, the H_0 for the Wilcoxon test is never rejected when the sample size is 5 and the significance level is 0.05. This leads to RP smaller than RP for Wilcoxon without preliminary test, and RP for t-test higher than RP for t-test without preliminary test.

As the sample size increases, RP for the t-test with a preliminary test decreases until it becomes slightly lower than RP for the t-test without a preliminary test in the non-rejection area. Meanwhile, RP for the Wilcoxon test, with and without the preliminary

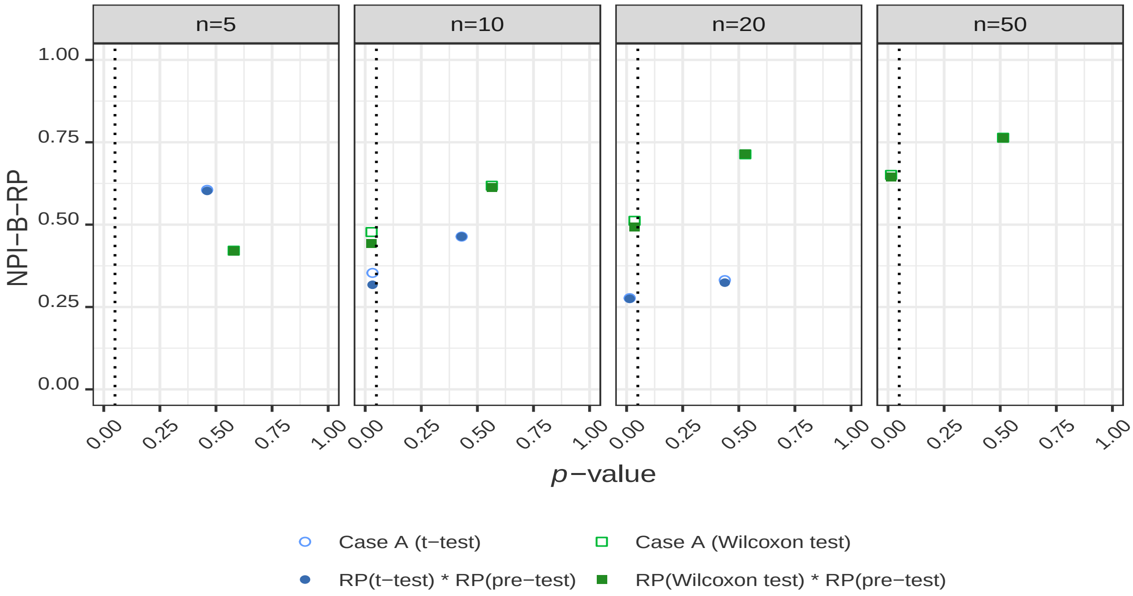


Figure 4.27: Comparing RP values for location tests with and without preliminary test (*Case A*), plotted against their corresponding mean p-values. When the original samples are drawn from $Ca(0, 1)$.

test, remains almost the same. This is because as the sample sizes increase, most NPI-B samples do not pass the Normality test and consequently perform the Wilcoxon test, whose power is slightly lower than that of the t-test with a large sample size. This leads to reduced RP for the t-test in *Case B* compared to RP for the t-test without a preliminary test and results in similar RP for the Wilcoxon test with and without a preliminary test.

Generally, with increasing sample size, the difference between the reproducibility of location tests with and without the preliminary test decreases. Thus, the preliminary Normality test does not substantially enhance the reproducibility of the location test outcomes.

4.4.3 The impact of the preliminary test on RP for location tests for *Case C*

Here the comparison results for RP for *Case C* and RP for location tests without preliminary tests are presented. We aim to determine whether filtering original samples based on a Normality test's outcome and applying a location test to all NPI-B samples without

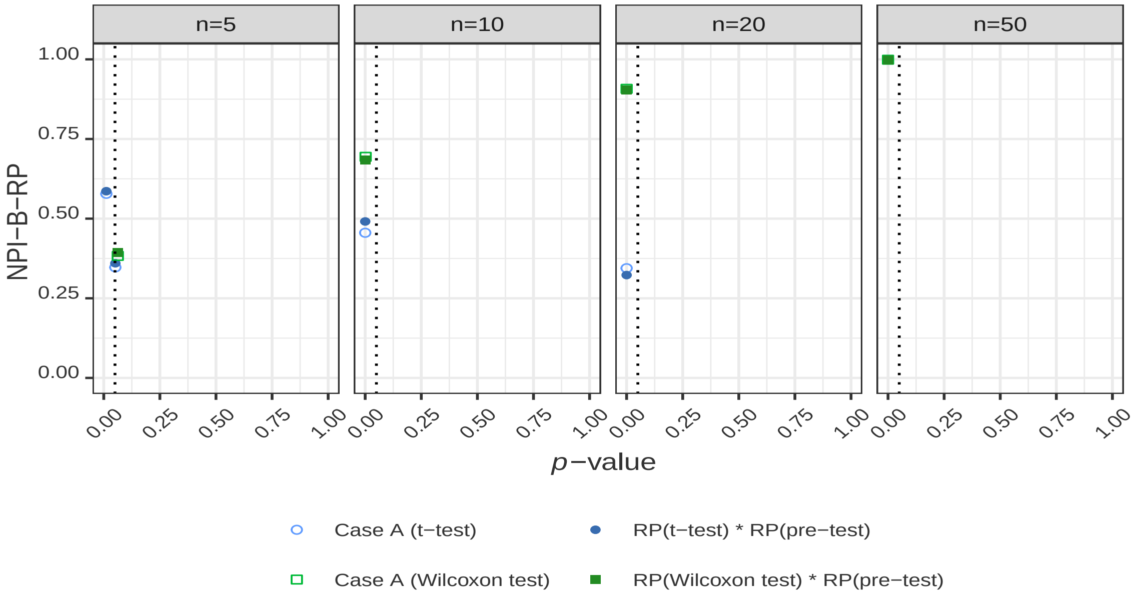


Figure 4.28: Comparing RP values for location tests with and without preliminary test (*Case A*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

further Normality testing can improve the RP of the location test results. By comparing the overall mean of RP values for the t-test after filtering the original samples based on passing the Normality test with the overall mean of RP values for the t-test without the Normality test for all original samples. In addition, comparing the overall mean of RP values for the Wilcoxon test after filtering the original samples based on failure to pass the Normality test with the overall mean of RP values for the Wilcoxon test without the Normality test for all original samples.

Figures 4.33 - 4.36 show this comparison for the investigated distributions and sample sizes. The impact of the preliminary test of Normality on the reproducibility of the location test results can vary depending on the distribution of the data and the type of location test being used. It appears that for the t-test, the impact of the preliminary Normality test is negligible when dealing with Normal distributions, as there is little difference in the overall mean RP values between *Case C* and the scenario without the Normality test. While, for the Wilcoxon test, there is slight variability in the RP values, indicating a small impact of the preliminary test. In contrast, when the data follow non-Normal distributions, the preliminary Normality test's impact on the Wilcoxon test's

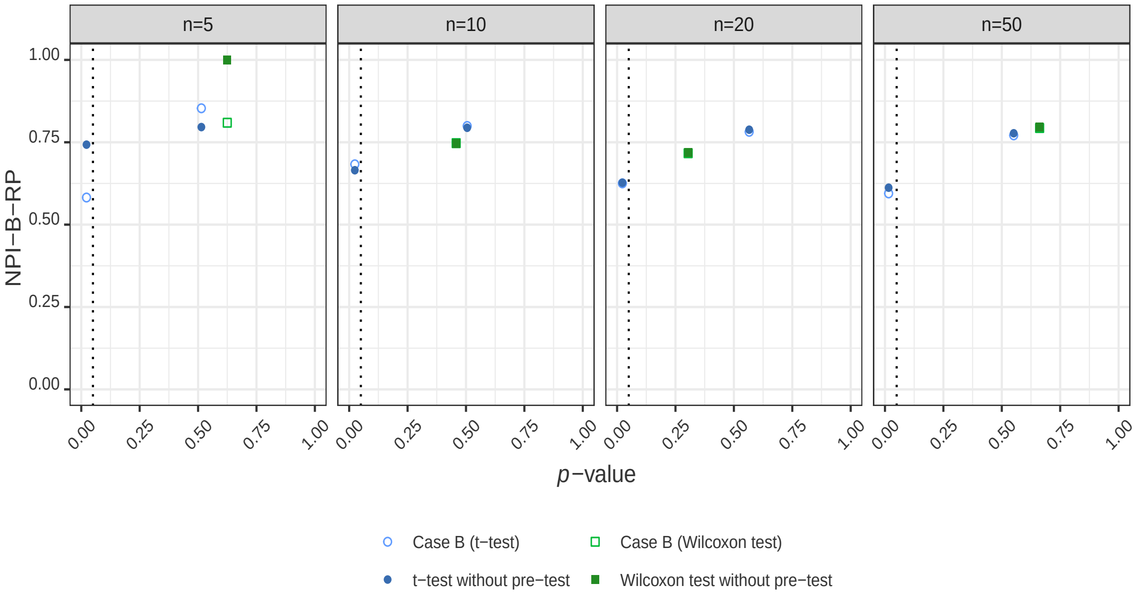


Figure 4.29: Comparing RP values for location tests with and without preliminary test (*Case B*), plotted against their corresponding mean p-values. When the original samples are drawn from $N(0, 1)$.

reproducibility is smaller than its impact on the t-test. This means that filtering original samples that meet the assumptions of the Normality test results leads to RP for the location tests very close to RP for the location tests without the Normality test. In general, this impact tends to be small. Thus, filtering the original samples according to the Normality test does not substantially improve the reproducibility of location tests.

The x-axis shows the overall mean of p-values for location tests after filtering original samples according to the Normality test and the overall mean of p-values for location tests without Normality tests for all original samples. Filtering original samples that meet the assumptions of the Normality test results leads to p-value close to the p-value for location tests without the Normality test.

These simulation studies were also conducted on other distributions under alternative hypotheses for location tests, and they led to results very similar to the results presented.

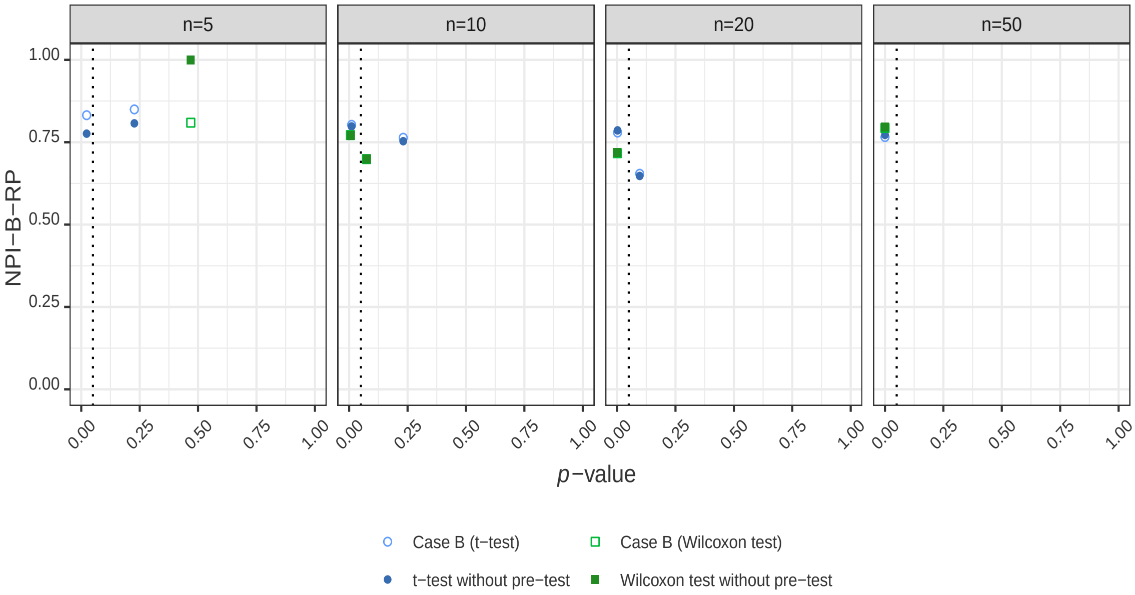


Figure 4.30: Comparing RP values for location tests with and without preliminary test (*Case B*), plotted against their corresponding mean p-values. When the original samples are drawn from $N(1, 1)$.

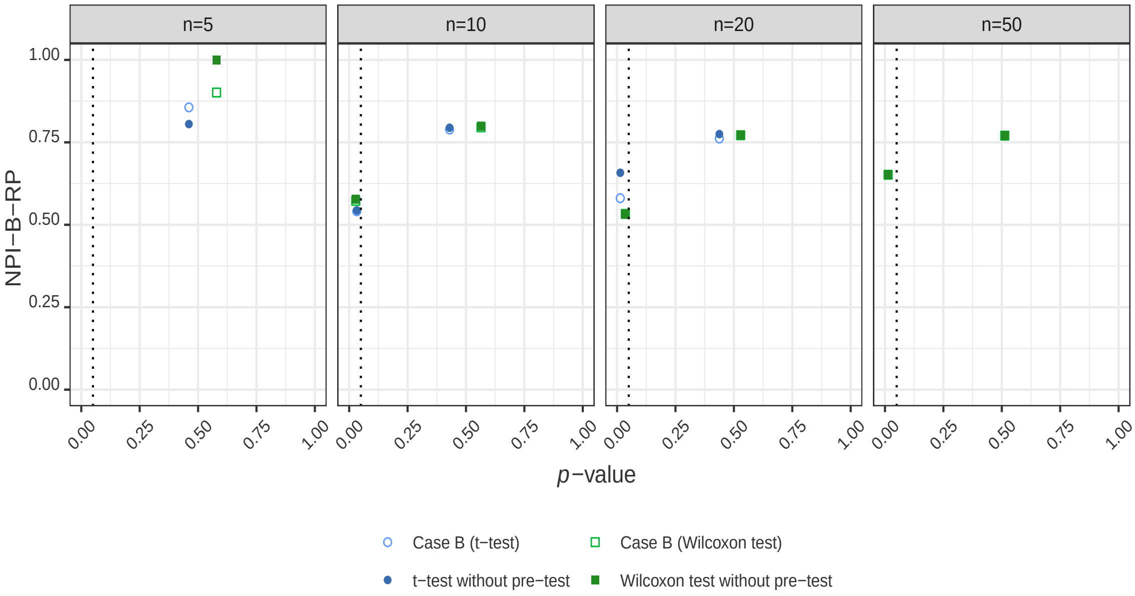


Figure 4.31: Comparing RP values for location tests with and without preliminary test (*Case B*), plotted against their corresponding mean p-values. When the original samples are drawn from $Ca(0, 1)$.

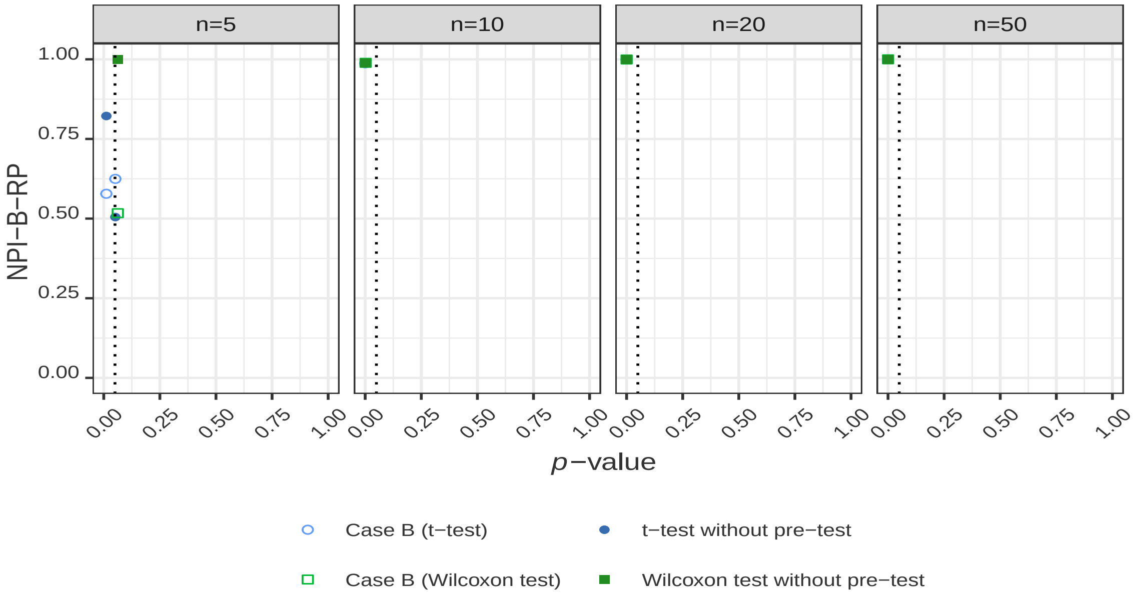


Figure 4.32: Comparing RP values for location tests with and without preliminary test (*Case B*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

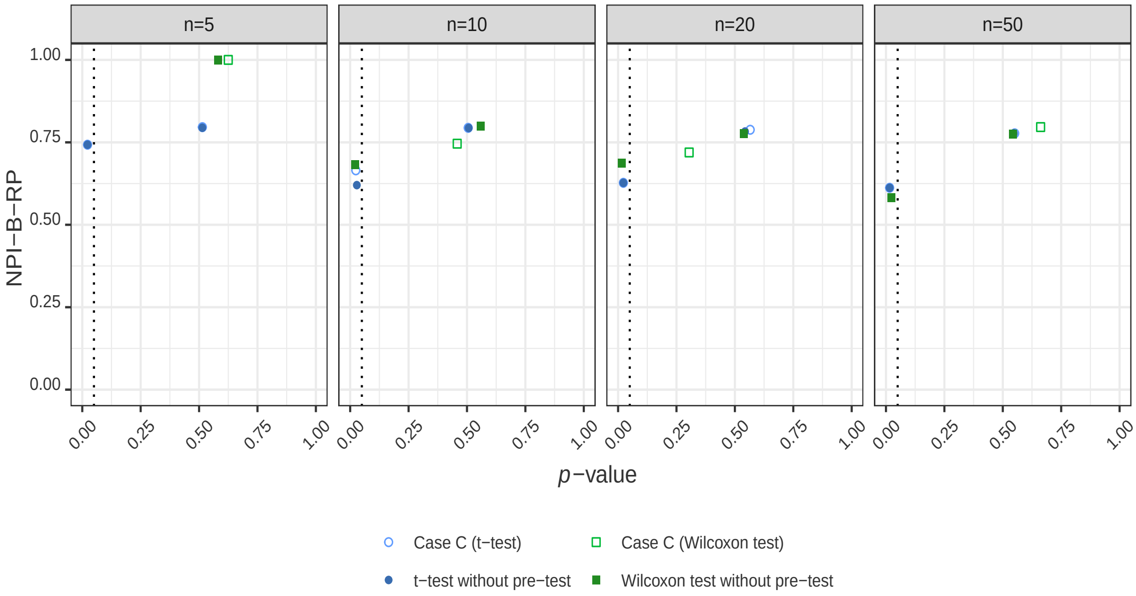


Figure 4.33: Comparing RP values for location tests with and without preliminary test (*Case C*), plotted against their corresponding mean p-values. When the original samples are drawn from $N(0, 1)$.

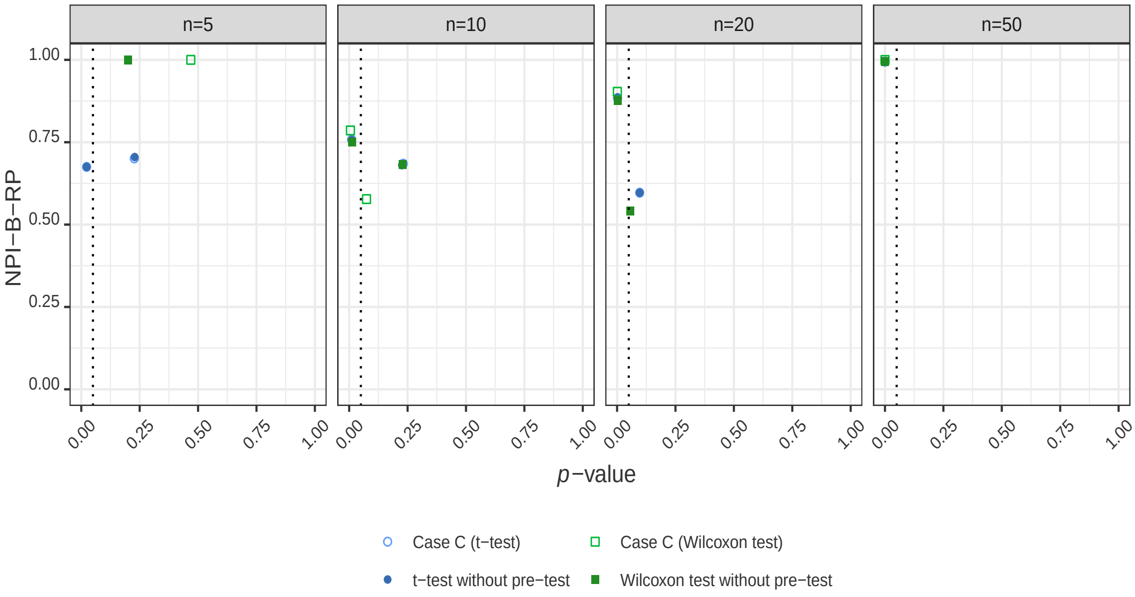


Figure 4.34: Comparing RP values for location tests with and without preliminary test (*Case C*), plotted against their corresponding mean *p*-values. When the original samples are drawn from $N(1,1)$.

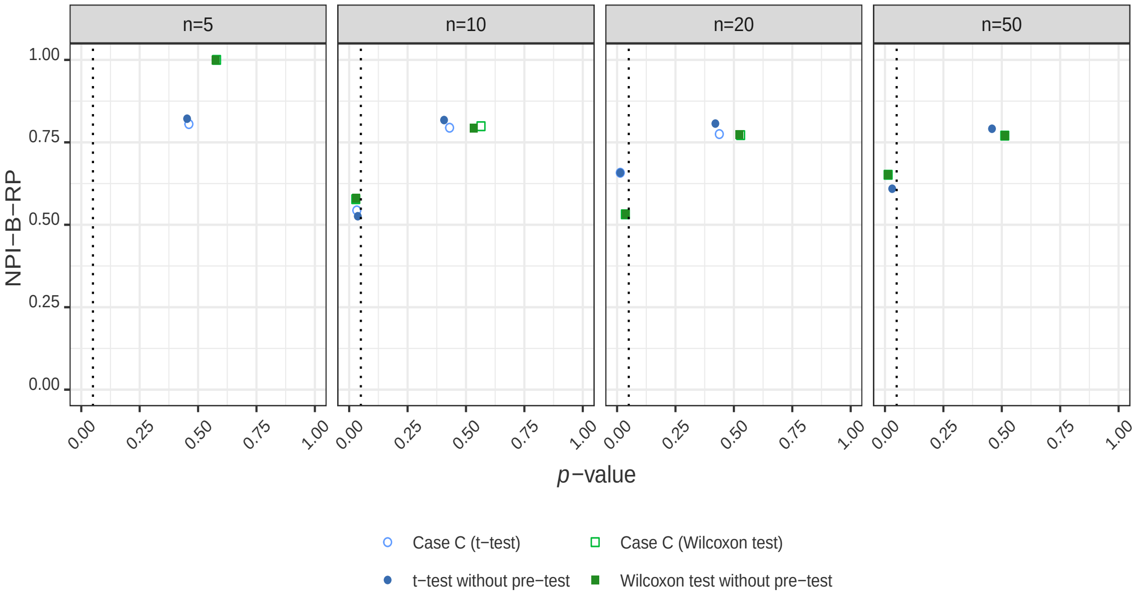


Figure 4.35: Comparing RP values for location tests with and without preliminary test (*Case C*), plotted against their corresponding mean *p*-values. When the original samples are drawn from $Ca(0,1)$.

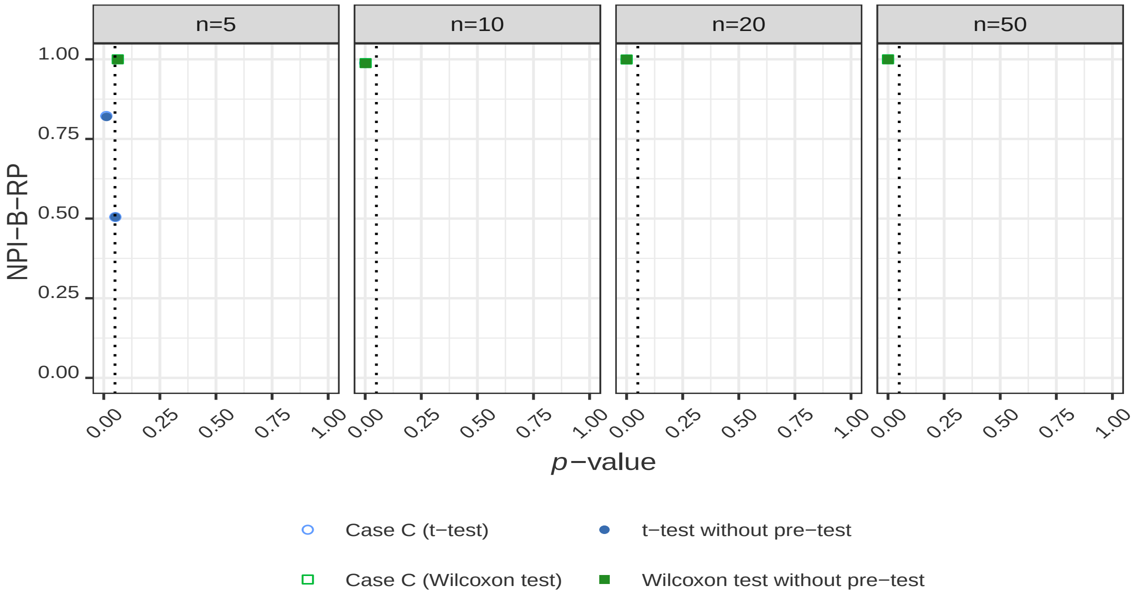


Figure 4.36: Comparing RP values for location tests with and without preliminary test (*Case C*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

4.5 Conclusions

This chapter explored the reproducibility probability (RP) for one-sample location tests, specifically focusing on the one-sample t-test and the one-sample Wilcoxon signed-rank test. The investigation covered RP for the location tests with and without the preliminary test for Normality through simulation studies. The RP values for the location tests with the preliminary test of Normality (the two-stage procedure) involve applying the Normality test in the first stage. Subsequently, if the null hypothesis for this Normality test is not rejected, then the t-test is performed in the second stage. Otherwise, the Wilcoxon test is applied. The objective is to assess the impact of the preliminary test on RP values for these location tests.

Three cases of reproducibility probability for the two-stage procedures were examined: *Case A* represents the full RP for the two-stage procedure. *Case B* represents the reproducibility of the same outcome for the location test, no matter which test is used. *Case C* represents reproducibility of the location test conclusion, where for the NPI-B samples the same location test is applied as for the original sample.

The results of the simulation studies show that the RP for the one-sample location test with and without preliminary test show the general pattern: as the p-value for the location tests is close to the threshold, the RP values are low. RP values are affected by the sample size, typically decreasing in the non-rejection area as the sample size increases, while increasing in the rejection area with larger sample sizes. Moreover, from the results of comparing the RP for the location tests with and without the preliminary test for Normality, there is no substantial difference between RP for the one-sample location test with and without the preliminary test of Normality for *Cases A, B* and *C*, which means that the effect of the preliminary test of Normality on RP of location test is small. This difference between the reproducibility of location tests with and without the preliminary test decreases as the sample size increases.

There is a relationship between the overall mean of RP values for the one-sample location tests with and without the preliminary test for Normality in the rejection area and their estimated power. As the power of the location tests increases the RP values in the rejection area increase. The power and the RP values in the rejection area for the location test increase as the sample size increases.

Chapter 5

Reproducibility of Two-Sample Location Tests with and without Preliminary Tests

5.1 Introduction

This chapter investigates the impact of preliminary tests on the reproducibility probability (RP) of two-sample location tests. The investigation considers the following three-stage procedure [78]: The Normality assumption is evaluated in the first stage. If the samples pass the Normality test, then the three-stage approach proceeds with a test for the equality of variances to determine if there is a significant difference between the variances of the two samples. If the null hypothesis for the equality of variances test is not rejected, then the two-sample t-test is applied. However, if the null hypothesis for the equality of variances test is rejected, then Welch's t-test is used. If one or both samples fail the Normality test, then the Wilcoxon-Mann-Whitney (WMW) test is applied. Additionally, the reproducibility of the two-sample location tests without applying preliminary tests is evaluated. Thus, the effect of the preliminary tests on RP of the two-sample location tests can be assessed by comparing RP for location tests with and without preliminary tests. Furthermore, the relationship between the mean of RP values in the rejection area with estimated power for the two-sample location tests with and without preliminary tests is examined.

This chapter is structured as follows: Section 5.2 provides a brief introduction to the two-samples location tests. Section 5.3 the reproducibility of the three-stage procedures for the two-sample location tests is investigated. Section 5.4 presents simulation studies to assess RP values for the three-stage procedure, as well as for location tests without preliminary tests. This section also provides the results of the simulations for RP of the location tests with and without preliminary tests, and displays the relationship between RP and the estimated power for location tests with and without preliminary tests. Comparison to identify the influence of the preliminary tests of Normality and equality of variances on RP values for the two-sample locations tests is presented in Section 5.5. Section 5.6 provides a summary of the key findings and results obtained in this chapter.

5.2 Two-sample location tests

This section provides a brief introduction to two-sample location tests, namely two-sample t-test, Welch's t-test, and Mann-Whitney U test.

5.2.1 Student's two-sample t-test

Suppose that a sample $\{X_1, X_2, \ldots, X_{n_X}\}$ from population X and an another sample $\{Y_1, Y_2, \ldots, Y_{n_Y}\}$ from population Y. The two-sample t-test is used to determine whether there is a statistically significant difference in the means of two independent samples, where the mean, denoted by μ, is a measure of central tendency that represents the average value of a dataset. The null hypothesis is $H_0 : \mu_X = \mu_Y$, and the alternative hypotheses are $H_1 : \mu_X \neq \mu_Y$ for two-sided testing. It is one widely used parametric test for this purpose first introduced by Gosset under the pen name Student [95]. Before conducting the t-test, certain assumptions need to be considered, including the homogeneity of variance and Normality of the data. The test statistic for Student's two-sample t-test is given by:

$$t = \frac{\bar{X} - \bar{Y}}{S_p\sqrt{\frac{1}{n_X} + \frac{1}{n_Y}}}$$

where $\bar{X}$ and $\bar{Y}$ are the means of samples, n_X and n_Y are the sample sizes, and S_p is the pooled standard deviation:

$$S_p = \sqrt{\frac{(n_X - 1)S_X^2 + (n_Y - 1)S_Y^2}{n_X + n_Y - 2}}$$

where S_X^2 and S_Y^2 represent the variances of samples. The null hypothesis is rejected in favour of the two-sided alternative if $|t| > t_{\alpha/2}$.

5.2.2 Welch's t-test

Welch t-test [101], also known as the unequal variance t-test or non-pooled variance t-test. It is a more robust alternative to Student's t-test when the assumption of equal population variances is not met or when sample sizes differ between groups [11]. The null hypothesis for the Welch's t-test is $H_0 : \mu_X = \mu_Y$, and the alternative hypothesis for the two-sided is $H_1 : \mu_X \neq \mu_Y$. Welch's t-test defines the test statistic t as follows [17]:

$$t = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\sqrt{(S_X^2/n_X) + (S_Y^2/n_Y)}} \tag{5.2.1}$$

Under H_0, Welch's t-test has a Student t-distribution with the number of degrees of freedom (df) derived using the Welch-Satterthwaite equation [17]:

$$df = \frac{\left(\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}\right)^2}{\left(\frac{S_X^2}{n_X}\right)^2 \left(\frac{1}{n_X - 1}\right) + \left(\frac{S_Y^2}{n_Y}\right)^2 \left(\frac{1}{n_Y - 1}\right)} \tag{5.2.2}$$

The null hypothesis is rejected when the p-value is less than or equal to the significance level α.

5.2.3 Mann-Whitney U test

The Mann-Whitney U test, also known as the Wilcoxon rank sum test or the Wilcoxon-Mann-Whitney (WMW) test, is a nonparametric test used as an alternative to the t-test for comparing two samples or groups when the data are not Normally distributed [42]. The Mann-Whitney test assesses whether the medians of the distributions for two populations are different from each other, where the median, denoted by η, is a measure of central tendency that represents the middle value of a dataset when arranged in ascending order. The null hypothesis is $H_0 : \eta_X = \eta_Y$, against the alternative hypothesis for two-sided testing $H_1 : \eta_X \neq \eta_Y$, where η_X is the median for population X, and η_Y is the median for population Y [65]. The observations from both populations are combined and ranked from smallest to largest. For each population, the ranks are summed separately. The

Mann-Whitney U test statistic denoted as U is calculated based on the rank sums [49]:

$$U = \min(U_X, U_Y)$$

where

$$U_X = n_X n_Y + \frac{n_X(n_X+1)}{2} - R_X$$
$$U_Y = n_X n_Y + \frac{n_Y(n_Y+1)}{2} - R_Y$$

where R_X is the sum of the ranks for population X, and R_Y is the sum of the ranks for population Y. If the null hypothesis holds, indicating that both sets of observations were obtained from identical populations, the median values would be similar and a significant overlap between the two samples would be expected [74]. The measure U represents this overlap between the two samples, with an expected similarity between U_X and U_Y [74]. Conversely, if there is minimal overlap, U will have significantly different values and $\min(U_X, U_Y)$ will be small [74]. Thus, if $U > U_{\text{critical}}$, we do not reject H_0. If $U \leq U_{\text{critical}}$, we reject H_0, where U_{critical} is the critical value of U obtained from the Mann-Whitney table [88].

5.3 Reproducibility of the three-stage procedure testing

In this section, the reproducibility from the NPI perspective for the three-stage procedure is evaluated. This procedure involves testing for Normality and homogeneity of variances for two groups. In the first stage, the Normality test is applied. If the null hypothesis of the Normality is rejected for one or both samples, then the Wilcoxon-Mann-Whitney test is used. Conversely, if both samples pass the Normality test, then the equality of variances is assessed using a preliminary test. If the null hypothesis for the equality of variances test is not rejected, then the two-sample t-test is applied, otherwise, Welch's t-test is used.

Three cases for studying the reproducibility of the three-stage procedure are considered, these are similar to the cases in Section 4.2.

1. *Case A*: Full reproducibility for the three stages. This case examines the RP for the same test outcome for all stages. For example, if original samples pass the Normality test and equality of variances test and H_0 for the t-test is not rejected, RP is computed as the ratio of the number of NPI-B samples that pass the Normality test and equality of variances test and H_0 for the t-test are not rejected to the total number of NPI-B samples.

2. *Case B*: the reproducibility for the same outcome for the location test, no matter which test is used. RP, in this case, is the ratio to obtain the same outcome for the location test (reject or not reject the null hypothesis) whether applying the t-test, Welch's t-test or WMW test across NPI-B samples similar to the outcome of the location test that applied in the original sample to the total number of NPI-B samples.

3. *Case C*: This case evaluates the RP of the location test conclusion, where for the bootstrap samples the same location test is applied as for the original sample without further preliminary testing. The RP is the ratio of NPI-B samples that have the same result (reject or not reject) for the null hypothesis of the location test that was applied to the original samples to the total number of NPI-B samples.

Suppose that N is the number of NPI-B samples and N_t^*, N_{Wt}^* and N_W^* are three disjoint subsets of $\{1, 2, \ldots, N\}$, $N_t^* \cap N_{Wt}^* \cap N_W^* = \emptyset$, where N_t^* is a subset of all the indices for the NPI-B samples that pass the Normality test and equality of variances test and t-test were applied, N_{Wt}^* is a subset of all the indices for the NPI-B samples that pass the Normality test but did not pass the equality of variances test and the Welch's t-test were performed, and N_W^* is a subset of all the indices for the NPI-B samples that do not pass the Normality test and the WMW test were applied. To compute NPI-B-RP for the three-stage procedure for two groups use the next steps:

Step 1: Perform a preliminary test for Normality separately on the original samples, with significance level α_1.

Step 2: If the null hypothesis H_0^N of the Normality test is not rejected, apply the preliminary test for equality variances and make a decision about its null hypothesis H_0^F with the level of significance α_2. If H_0^F is not rejected, then apply t-test with α_3 and decide H_0^t, set $TS^t = 1$ if H_0^t is rejected or set $TS^t = 0$ if H_0^t is not rejected. If H_0^F is rejected, then apply Welch's t-test with α_3 and decide H_0^{Wt}, set $TS^{Wt} = 1$ if H_0^{Wt} is rejected or set $TS^{Wt} = 0$ if H_0^{Wt} is not rejected. Whereas, if H_0^N for both or one sample is rejected, then apply the WMW test with the level of significance α_3 and decide H_0^W, set $TS^W = 1$ if H_0^W is rejected or set $TS^W = 0$ if H_0^W is not rejected.

Step 3: Draw an NPI-B sample N times for each of the original samples, the same size as the original samples.

- For *Cases A* and *B*, perform the Normality test and equality of variances test as preliminary tests on the N pair of NPI-B samples, then t-test, Welch's t-test, or Wilcoxon test according to test decision of the preliminary tests. Each time record the test decision $TS_i^t = 1$ if H_0^t is rejected, $TS_f^{Wt} = 1$ if H_0^{Wt} is rejected, or $TS_j^W = 1$ if H_0^W is rejected, or record $TS_i^t = 0$ if H_0^t is not rejected, $TS_f^{Wt} = 0$ if H_0^{Wt} is not rejected, or $TS_j^W = 0$ if H_0^W is not rejected, where $i \in N_t^*$, $f \in N_{Wt}^*$ and $j \in N_W^*$.
- For *Case C*, the same location test is performed on the N pair of NPI-B samples as for the original samples. Each time record the test decision $T_s = 1$ if H_0^t is rejected, or $WT_s = 1$ if H_0^{Wt} is rejected, or $W_s = 1$ if H_0^W is rejected, or record $T_s = 0$ if H_0^t is not rejected, or $WT_s = 0$ if H_0^{Wt} is not rejected, or $W_s = 0$ if H_0^W is not rejected, where $s = 1, \ldots, N$

Step 4: Compute the RP based on the test decisions of the NPI-B samples.

(i) The RP for *Case A*:

If both original samples passed the Normality test and the equality of variances test and t-test was applied the RP is

$$RP_t = \sum_{i \in N_t^*} \mathbb{I}_{\{TS^t = TS_i^t\}} \frac{1}{N}$$

If both original samples passed the Normality test and did not pass the equality of variances test and Welch's t-test was applied the RP is

$$RP_{Wt} = \sum_{f \in N_{Wt}^*} \mathbb{I}_{\{TS^{Wt} = TS_f^{Wt}\}} \frac{1}{N}$$

If one or both original samples did not pass the Normality test and the Wilcoxon test was applied the RP is

$$RP_W = \sum_{j \in N_W^*} \mathbb{I}_{\{TS^W = TS_j^W\}} \frac{1}{N}$$

(ii) The RP for *Case B*:

If both original samples passed the Normality test and the equality of variances test, and t-test was applied the RP is

$$RP_t = \left(\sum_{i \in N_t^*} \mathbb{I}_{\{TS^t = TS_i^t\}} + \sum_{f \in N_{Wt}^*} \mathbb{I}_{\{TS^t = TS_f^{Wt}\}} + \sum_{j \in N_W^*} \mathbb{I}_{\{TS^t = TS_j^W\}} \right) \frac{1}{N}$$

If both original samples passed the Normality test and did not pass the equality of variances test and Welch's t-test was applied the RP is

$$RP_{Wt} = \left(\sum_{i \in N_t^*} \mathbb{I}_{\{TS^{Wt} = TS_i^t\}} + \sum_{f \in N_{Wt}^*} \mathbb{I}_{\{TS^{Wt} = TS_f^{Wt}\}} + \sum_{j \in N_W^*} \mathbb{I}_{\{TS^{Wt} = TS_j^W\}} \right) \frac{1}{N}$$

If one or both original samples did not pass the Normality test and the Wilcoxon test was applied the RP is

$$RP_W = \left(\sum_{i \in N_t^*} \mathbb{I}_{\{TS^W = TS_i^t\}} + \sum_{f \in N_{Wt}^*} \mathbb{I}_{\{TS^W = TS_f^{Wt}\}} + \sum_{j \in N_W^*} \mathbb{I}_{\{TS^W = TS_j^W\}} \right) \frac{1}{N}$$

(iii) The RP for *Case C*:

If both original samples passed the Normality and the equality of variances tests and t-test was applied the RP is

$$RP_t = \sum_{s=1}^{N} \mathbb{I}_{\{TS^t = T_s\}} \frac{1}{N}$$

If both original samples passed the Normality test but did not pass the test for equality of variances and Welch's t-test was applied the RP is

$$RP_{Wt} = \sum_{s=1}^{N} \mathbb{I}_{\{TS^{Wt} = WT_s\}} \frac{1}{N}$$

If one or both original samples did not pass the Normality test and the Wilcoxon test was applied the RP is

$$RP_W = \sum_{s=1}^{N} \mathbb{I}_{\{TS^W = W_s\}} \frac{1}{N}$$

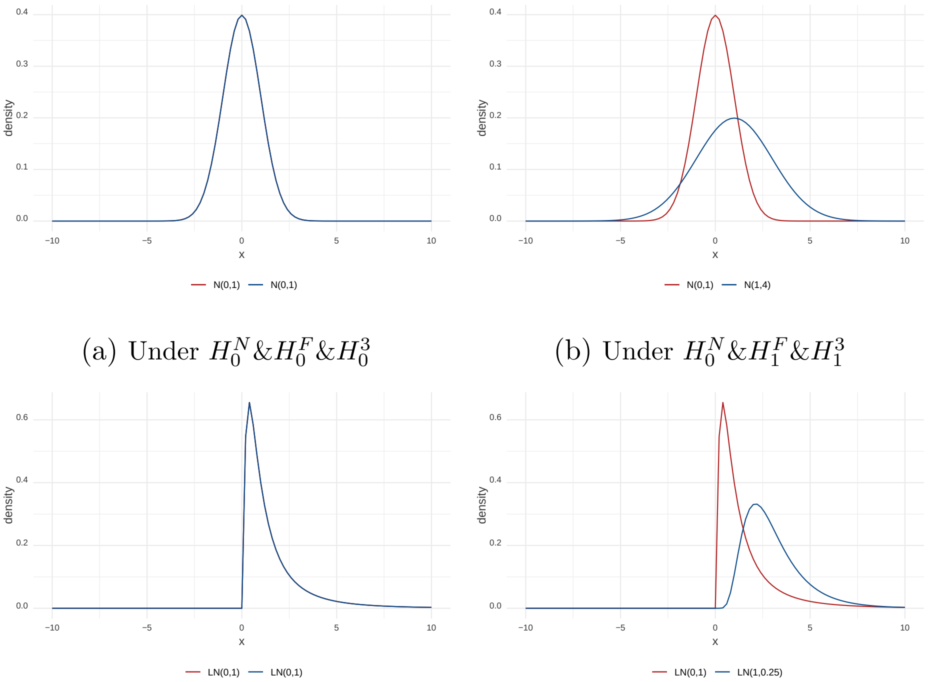
Step 5: Perform Steps 3 and 4 in total h times, record the outcomes by RP_{t_k},RP_{Wt_k}, or RP_{W_k}, where $k = 1, 2, \ldots, h$.

5.4 Simulation studies for the reproducibility of the location tests with and without preliminary tests

Simulation studies are performed to investigate the reproducibility of the three-stage procedure including testing for Normality and homogeneity of the variances and location tests. We considered three main statistical tests, the t-test, Welch's t-test and the Wilcoxon-Mann-Whitney test. For the preliminary tests, we used the Shapiro-Wilk test for testing the Normality which is independently conducted for each sample and the F-test for testing the homogeneity of the variances of the two samples. These preliminary tests are chosen because the Shapiro-Wilk test is known to be sensitive to deviations from Normality, especially with small to moderate sample sizes. It gives high reproducibility and can be applied to a sample size as small as 5, this was addressed in Chapter 2. The F-test is easy to use; it provides a simple and straightforward way to determine if the variances are significantly different. It also gives good reproducibility when the assumption of Normality is met, its reproducibility was studied in Chapter 3. We apply these simulations by performing the Steps in Section 5.3 with the inputs $N = 1000$ and $h = 100$, and choosing the minimum, mean, and maximum from these h RP values.

Moreover, simulation studies are conducted to investigate reproducibility for the two-sample location tests without the preliminary test for Normality or equality of variances. This simulation is carried out by applying the NPI-B-RP Algorithm 1, as presented in Section 1.4.5 of Chapter 1.

The null hypothesis for the preliminary Normality test is H_0^N : the population is Normally distributed, against the alternative hypothesis is H_1^N : the population is not Normally distributed. Whereas the null hypothesis for the preliminary equality of variances test is $H_0^F : \sigma_X^2 = \sigma_Y^2$, and the alternative hypothesis is $H_1^F : \sigma_X^2 \neq \sigma_Y^2$, where σ_X^2 and σ_Y^2 are the populations variances. For the location tests, the null hypothesis is $H_0^3 : \theta_X = \theta_Y$, where θ_X and θ_Y are the location parameter for populations X and Y, respectively, H_0^3 denotes the null hypothesis for the third stage of the location test which is either t-test (H_0^t) hypothesis or Welch's t-test (H_0^{Wt}) hypothesis in which case θ is the mean, or the null hypothesis for WMW test (H_0^W) in which case θ is the median. The



(a) Under $H_0^N \& H_0^F \& H_0^3$

(b) Under $H_0^N \& H_1^F \& H_1^3$

(c) Under $H_1^N \& H_0^F \& H_0^3$

(d) Under $H_1^N \& H_1^F \& H_1^3$

Figure 5.1: PDFs for the chosen distributions for two groups used in the simulation studies.

corresponding alternative hypothesis is $H_1^3 : \theta_X \neq \theta_Y$.

To simulate data for the three-stage procedure, different distributions are considered, and their probability density functions are shown in Figure 5.1. We simulate data under H_0^N, H_0^F and H_0^3, we generate data from $N(0,1)$ for pair original samples. Also, data are simulated under H_0^N, H_1^F and H_1^3, we choose $N(0,1)$ and $N(1,2^2)$ distributions. In addition, we chose Log Normal distribution $LN(0,1)$ for both samples to study RP under H_1^N, H_0^F and H_0^3. Finally, we simulate data under H_1^N, H_1^F and H_1^3, where data are created from Log Normal distribution $LN(0,1)$ and $LN(1,0.5)$.

The number of runs per simulation is $K = 100$, for each run, two original samples of size n are generated from the chosen distributions and perform the steps in Section 5.3. Different sample sizes $n_X = n_Y = 5, 10, 20, 50$ are considered. All tests considered are two-sided tests with a significance level of 0.05.

5.4.1 The results of the reproducibility for three-stage procedure

This part presents the results of RP for the three-stage procedure that is RP for the location with the preliminary tests for Normality and equality of variances for two groups for three *Cases A, B,* and *C.*

The colour scheme for the plots in this thesis is designed to provide clear visualization, facilitating interpretation of the results. The blue colour is reserved for parametric tests which are used when the assumptions of Normality and equality of variances are met. The red colour denotes Welch's parametric tests, which are used when the assumption of equality of variances is violated. Finally, the green colour represents nonparametric tests which do not make assumptions about the distribution of the data.

The results for *Case A*

The results of the simulation for the full reproducibility of the three stages for *Case A* are shown in Figures 5.2 - 5.5. The RP values show the general pattern: RPs tend to be lower when the p-value is close to the significance threshold (α), and RPs increase as the p-value moves further away from the threshold. However, there is extreme variability in the RP values for the location tests, especially the parametric tests.

The influence of the preliminary tests for Normality and equality of variances is notably evident in the analysis of original samples, particularly when dealing with large sample sizes. For instance, in scenarios where the distributions are Normal and exhibit equal variances, the majority of original samples apply the t-test, as shown in Figure 5.2. If the distributions are Normal but have different variances, then most original samples choose Welch's t-test, as shown in Figure 5.3. In cases where the distributions are non-Normal, the prevailing choice for the majority of original samples is the WMW test, clearly demonstrated in Figures 5.4 and 5.5.

The RP values for the WMW test tend to increase as the sample size increases. Conversely, RP values for the parametric tests decrease, as the sample size increases. This is due to the Normality test ability; for the large sample size, the ability of the Normality

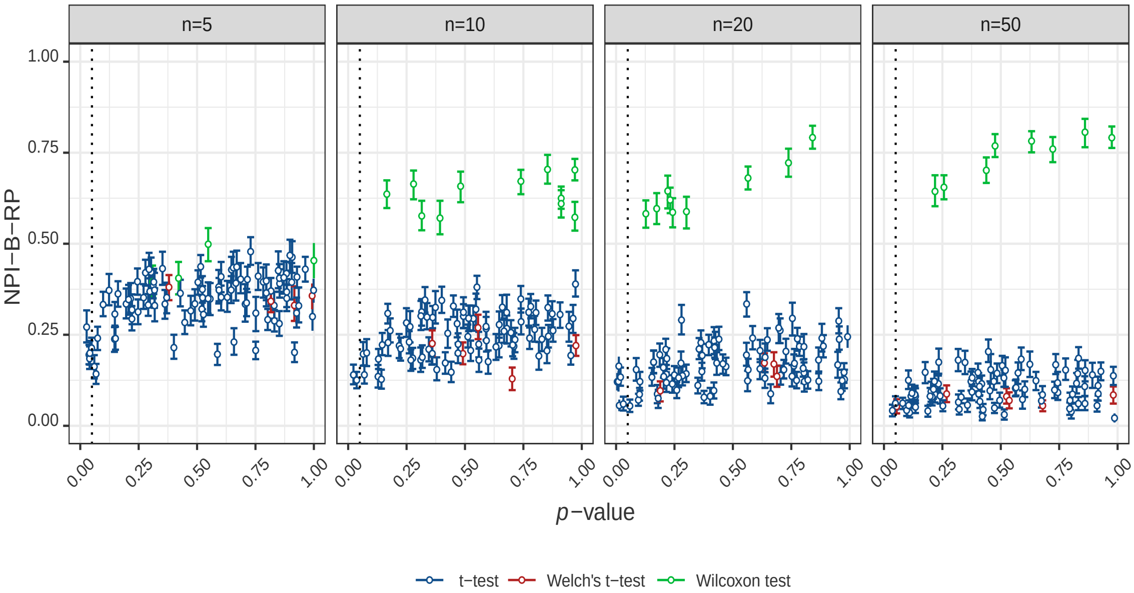


Figure 5.2: The min, mean, and max of RP values for the three-stage procedure against p-values for the location test stage. Original samples were drawn from $N(0,1)$, *Case A*.

test to detect skewness in the various NPI-B samples becomes strong, which leads to non-exceeding a lot of NPI-B samples the Normality test and performing the WMW test, resulting in high RP values for the WMW test and low RP for parametric tests and vice versa.

When the variances of the distributions are equal, the RP for the t-test is slightly better than those for Welch's t-test, as shown in Figure 5.2. Conversely, in cases of unequal variances, the RP for Welch's t-test tends to be a little bit better than that for those for the t-test, as shown in Figure 5.3. When the sample size is large and the distributions deviate more from Normality, the RP values for the WMW test with the preliminary test of Normality are high.

The results for *Case B*

This part shows the simulation results for RP values for the same outcome for the location tests, no matter which test is used. Figures 5.6 - 5.9 illustrate these RP values obtained from the simulations, these RP values show the general pattern. The RP values of the location tests are affected by sample size, RP values tend to decrease in the non-rejection area, as the sample size increases. While in the rejection area, RP values increase as the

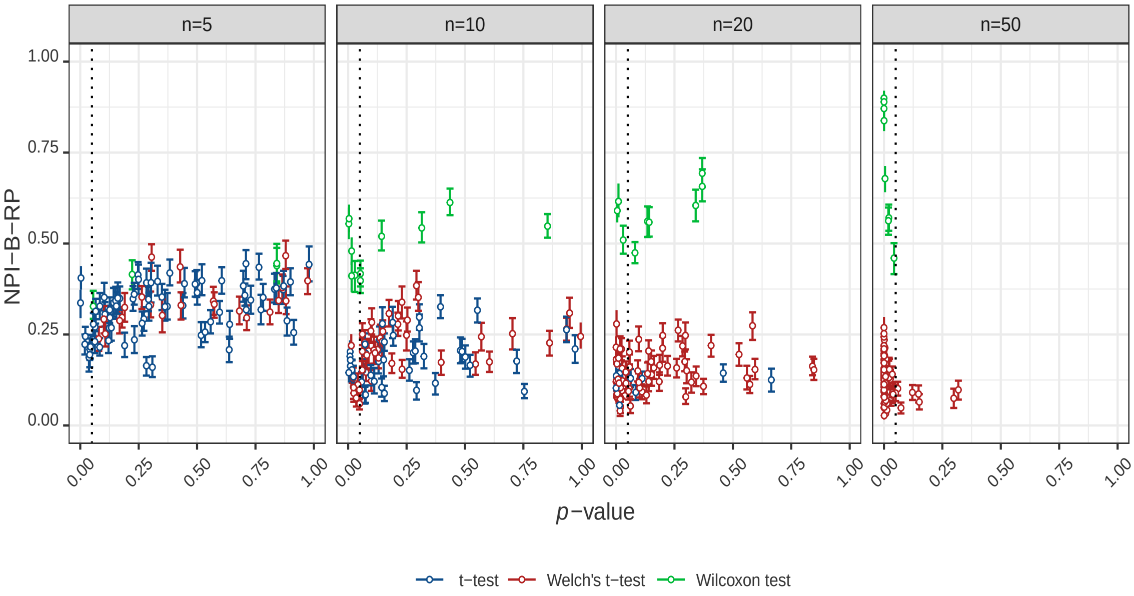


Figure 5.3: The min, mean, and max of RP values for the three-stage procedure against p-values for location test. Original samples were drawn from $N(0,1)$ and $N(1,2^2)$, *Case A*.

sample size increases.

When original samples are drawn from $N(0,1)$ and $N(1,2^2)$, there is slight variability in RP values for Welch's t-test in the non-rejection area shown in Figure 5.7. This is because RP for *Case B* depends on NPI-B samples having the same outcome as the original sample, whether performing the t-test Welch's t-test or the WMW test. For large sample sizes in the non-rejection area, most NPI-B samples do not pass the Normality test, resulting in the use of the WMW test. The RP for the WMW test exhibits variability when the variances are different, which in turn affects the RP for original samples that perform Welch's t-test.

The results for *Case C*

The results of the simulation for the reproducibility of the three-stage procedure for *Case C* are presented in Figures 5.10 - 5.13. The RP values for location tests with preliminary tests follow the general pattern: RP is low when the p-value is close to the threshold and high when the p-value is far away from the threshold. The results are similar to those observed in *Case B*, however, the variability observed in RP values for Welch's t-test in *Case B* in Figure 5.7 decrease in *Case C* as shown in Figure 5.11 because RP for this

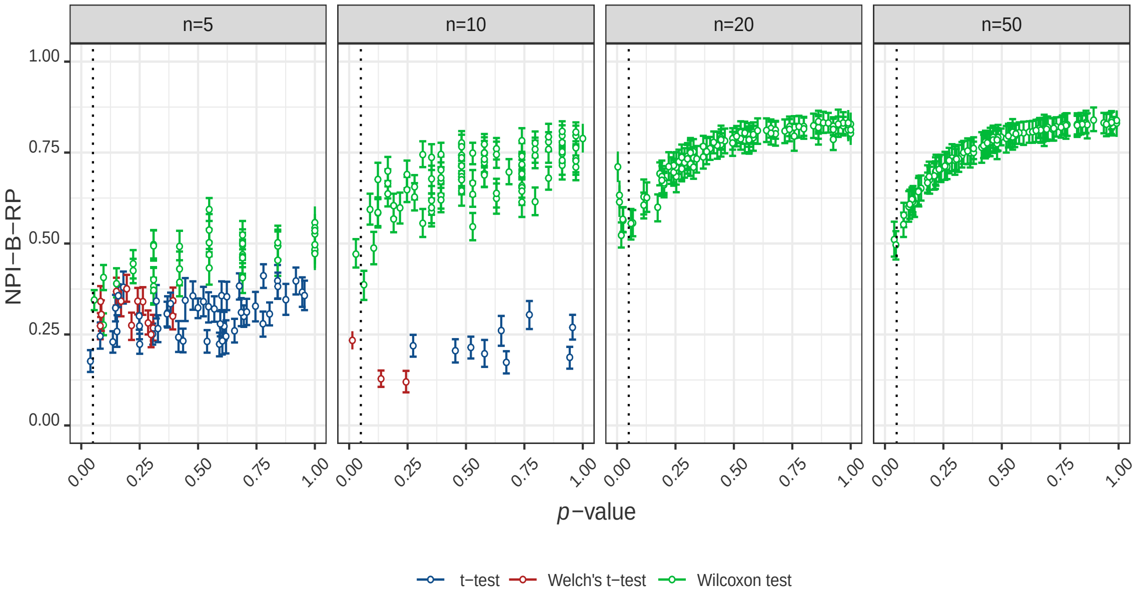


Figure 5.4: The min, mean, and max of RP values for the three-stage procedure against p-values for the location test stage. Original samples were drawn from $LN(0,1)$, *Case A*.

Sample size	RP	Power
$n = 5$	0.553	0.143
$n = 10$	0.581	0.269
$n = 20$	0.668	0.491
$n = 50$	0.778	0.872

Table 5.1: The relationship between RP and power for the three-stage procedure, samples from $N(0,1)$ and $N(1,2^2)$.

case depends only on NPI-B samples that perform the same test as the original sample.

The relationship between RP and the estimated power for the three-stage procedure

The relationship between the overall mean of RP values in the rejection area and the estimated power for the three-stage procedure under the alternative hypothesis for the location tests is examined. The overall mean of RP values for samples in the rejection area out of 100 original samples. Where the three-stage testing procedure applies the Shapiro-Wilk (SW) test and F-test to choose between location test t-test, Welch's t-test,

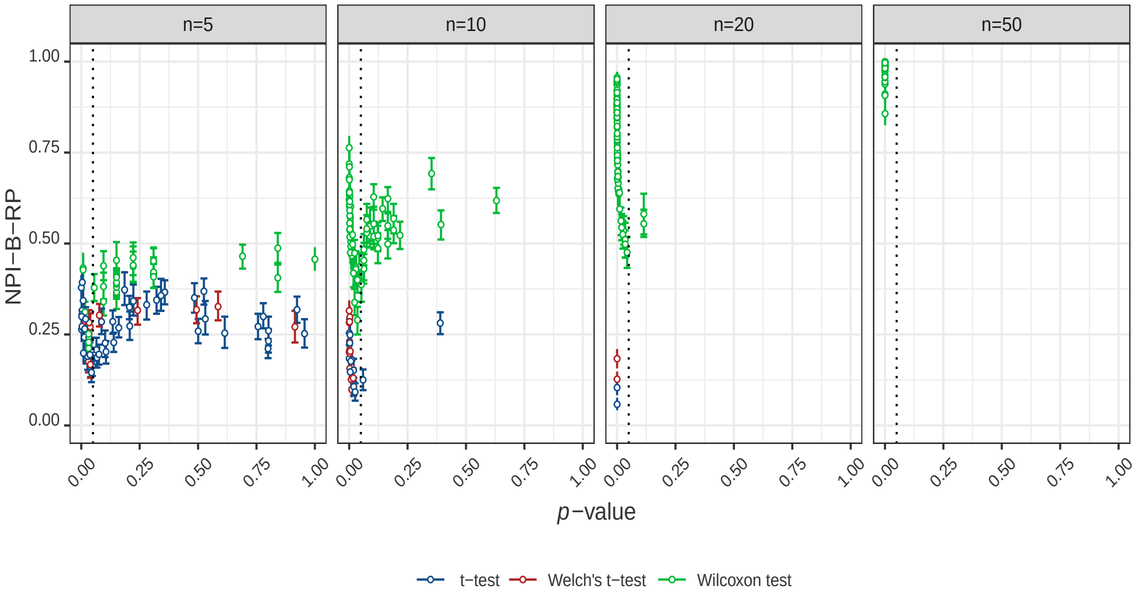


Figure 5.5: The min, mean, and max of RP for the three-stage procedure against p-values for location test. Original samples were drawn from $LN(0,1)$ and $LN(1,0.5^2)$, *Case A*.

Sample size	RP	Power
$n = 5$	0.599	0.354
$n = 10$	0.726	0.714
$n = 20$	0.960	0.960
$n = 50$	0.987	1.000

Table 5.2: The relationship between RP and power for the three-stage procedure, samples from $LN(0,1)$ and $LN(1,0.5)$.

or WMW test on both original sample and NPI-B samples. RP represents the proportion of NPI-B samples that get the same location test result as the original sample to the total number of NPI-B samples, regardless of which location test is used.

A Monte Carlo simulation of $10,000$ datasets is performed to estimate the power of the three-stage testing procedure, where the SW test for Normality and F-test are conducted as preliminary tests and then proceed with either the t-test, Welch's t-test or the Wilcoxon test based on the outcome of the preliminary tests. The power will be the proportion of datasets for which of the tests successfully rejects the null hypothesis H_0^2 to the total number of datasets.

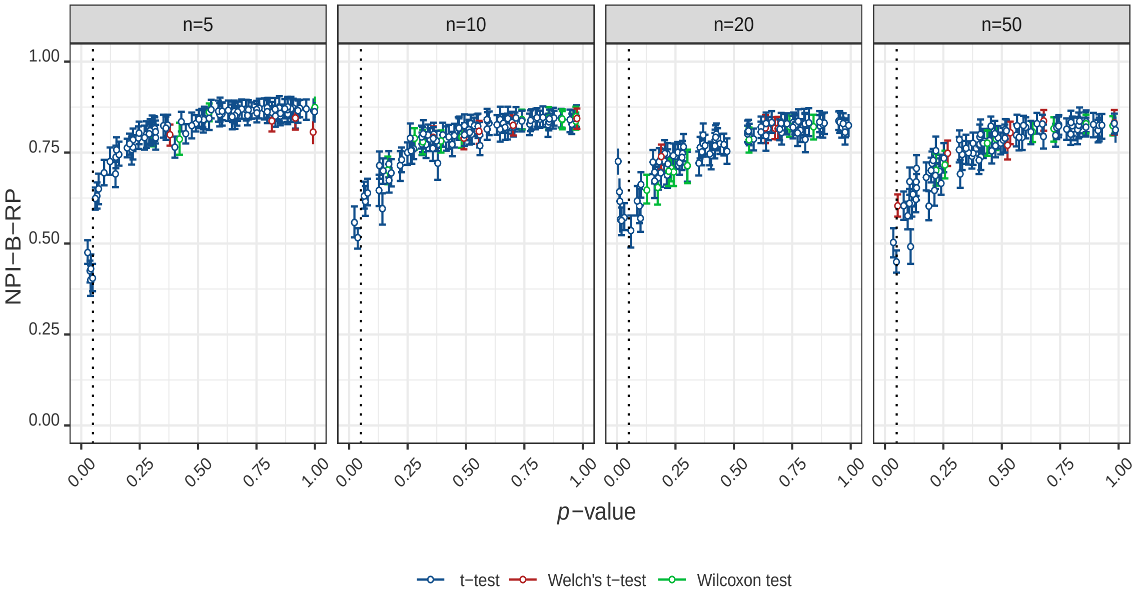


Figure 5.6: The min, mean, and max of RP values for the three-stage procedure *Case B* against p-values for the location test stage, samples are drawn from $N(0, 1)$.

We observed from Table 5.1 that shows this relationship under the alternative hypothesis for the location tests for Normal distributions and from Table 5.2 that shows non-Normal distributions, the RP for location tests with preliminary tests increases as their power increase. Additionally, power and RP increase with increasing sample size.

5.4.2 The results of the reproducibility for the location tests without preliminary tests

This section shows the findings derived from simulations of location tests (the two-sample t-test, Welch's t-test, and WMW test) conducted without preliminary tests. The RP values for each test were presented in separate figures in Appendix D.2. Here, we will be limited to displaying the mean of RP values for each test in one plot.

Figure 5.14 shows the results of simulations for the RP values for the location tests without the preliminary tests. When both original samples are drawn from Normal distributions having identical mean values and equality variances $N(0, 1)$ (where $\mu_X = \mu_Y = 0$, and $\sigma_X = \sigma_Y = 1$). The vast majority of original samples are located in the non-rejection area since both original samples are drawn from distributions that have the same means.

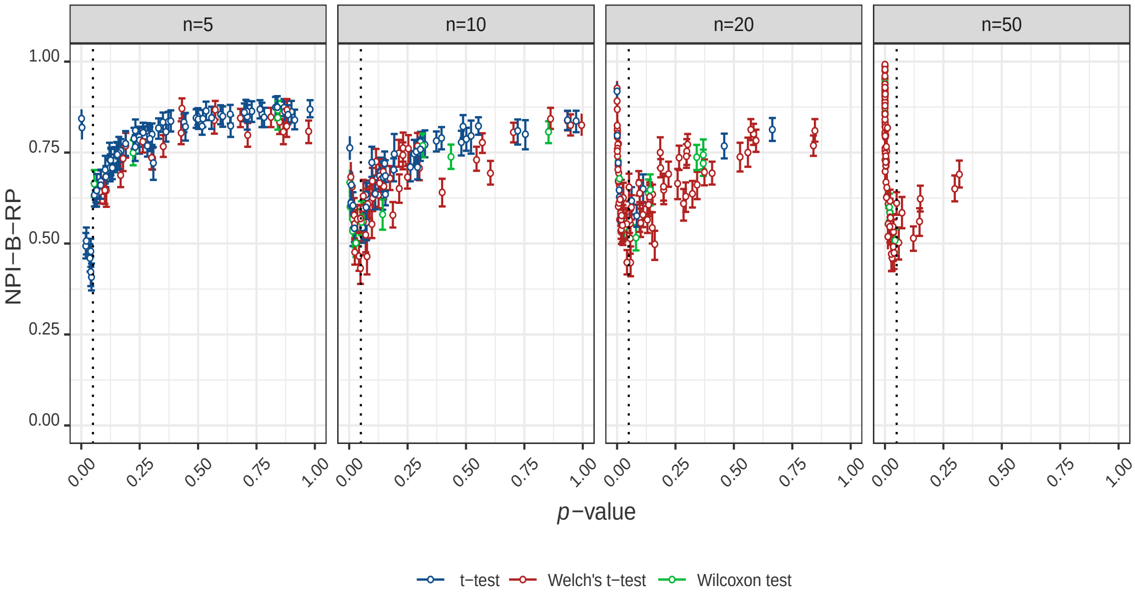


Figure 5.7: The min, mean, and max of RP values for the three-stage procedure *Case B* against p-values for the location test, samples are drawn from $N(0,1)$ and $N(1,2^2)$.

The RP values for location tests show the general pattern: RP increases gradually as their p-values move away from the threshold $\alpha = 0.05$. Also, RP for the location test does not reach close to 1 when the p-value is close to 1. This is because the tests are conducted for the two-sided when the p-value close to 1 indicates that the evidence provided by the data is not strong enough to reject H_0, but it also does not strongly support it either. This uncertainty can lead to different test outcomes, reducing the reproducibility of the results.

When comparing the RP for these location tests: There exists noticeable variability in RP for location tests when dealing with small sample sizes. However, as the sample size increases, this variability decreases. This is because, for small sample sizes, the power of the location test tends to be low. This lower power results in more diverse results, leading to greater variability in RP. As the sample size increases, the power of location tests increases because larger sample sizes provide more information, leading to more precise estimates of population parameters, resulting in less variability in RP.

Moreover, RPs for the WMW test seem slightly higher than RPs for Welch's t-test followed by RPs for the t-test for the small sample size $n = 5$ in the non-rejection area. As the sample size increases, RPs for the WMW test decrease slightly than those for the

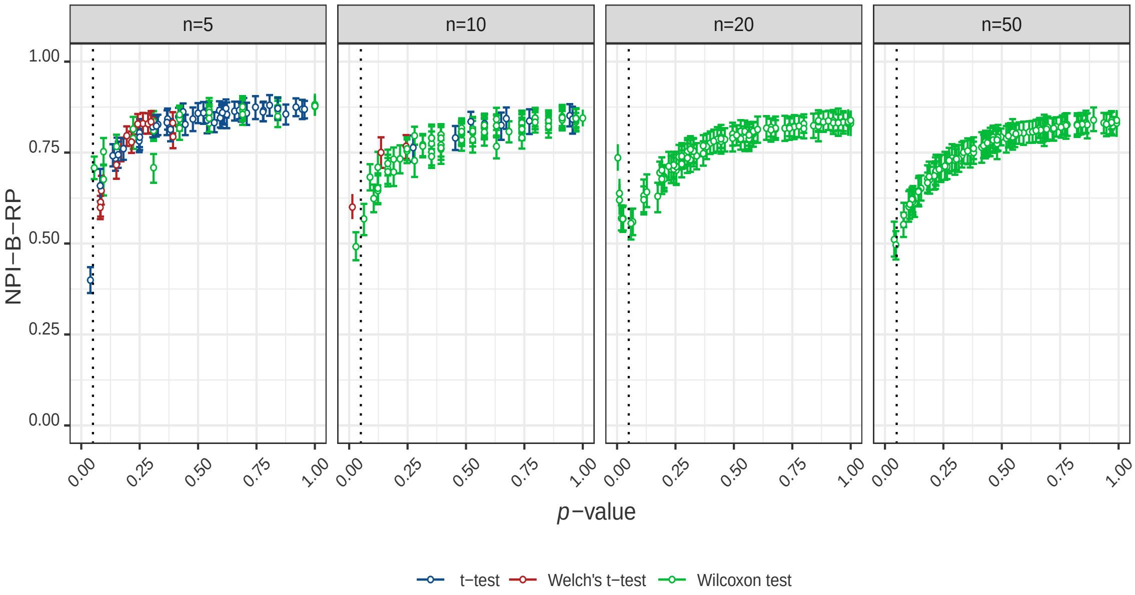


Figure 5.8: The min, mean, and max of RP values for the three-stage procedure *Case B* against p-values for the location test stage, samples are drawn from $LN(0,1)$.

t-test and Welch's t-test. This is because, for smaller sample sizes, it is more likely that the assumptions of Normality and equal variances will be violated. This can impact the effectiveness of parametric tests (t-test and Welch's t-test), leading to slightly lower RPs compared to the WMW test. However, with larger sample sizes, parametric tests become less sensitive to violations of these assumptions, resulting in an improvement in their RP. On the other hand, because the WMW test does not depend on these assumptions, its RPs may not show change with increasing sample size. In addition, as the sample size increases RPs for the t-test and Welch's t-test become similar. This is because the impact of unequal variances and non-Normality decreases with larger sample sizes, and the t-test remains robust even when the equal variance assumption is violated.

Figure 5.15 presents simulation results for the RP of location tests without the preliminary tests. The simulation was conducted with original samples drawn from Normal distributions $N(0,1)$ and $N(1,2^2)$ (where $\mu_X = 0, \mu_Y = 1$ and $\sigma^2_X = 1, \sigma^2_Y = 4$). For small sample sizes, there are many original samples in the non-rejection area (where tests fail to detect significant differences). However, as the sample size increases, most of the original samples shift to the rejection area (where tests detect significant differences) with high

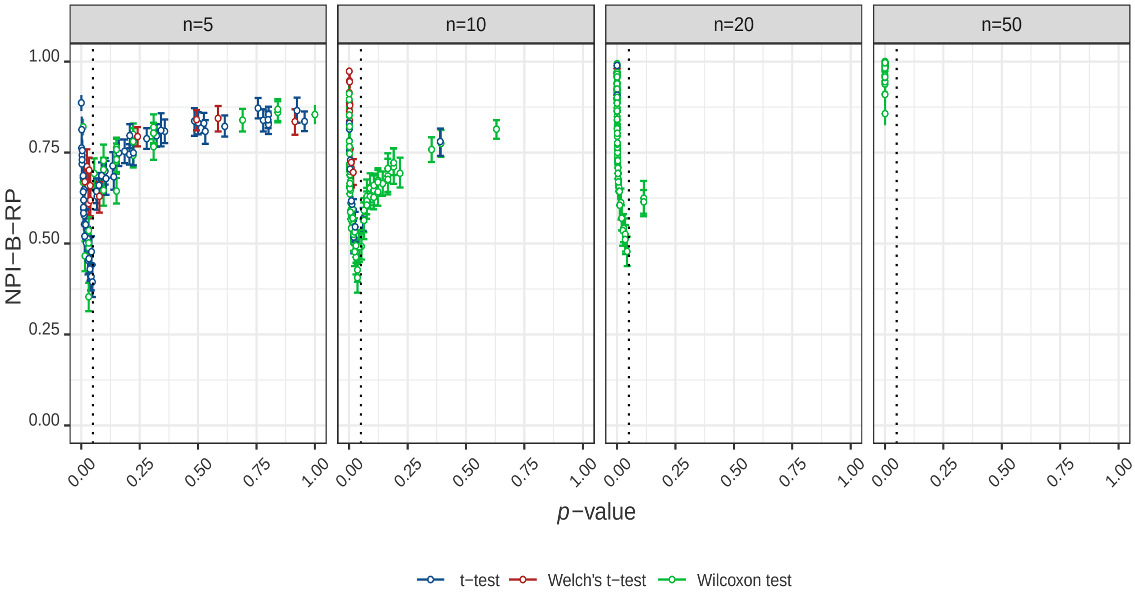


Figure 5.9: The min, mean, and max of RP for the three-stage procedure *Case B* against p-values for the location test, samples are drawn from $LN(0,1)$ and $LN(1,0.5)$.

RP values. Also, the p-values for the location test become closer to the threshold in the non-rejection area with low RP values, as the sample size increases. While in the rejection area, the RP values increase until they reach one when the p-value decreases to zero. This is because the power of the location tests increases as the sample size increases, and a greater proportion of original samples and the NPI-B samples show significant differences since the means of distributions are different, resulting in higher reproducibility in the rejection area and lower reproducibility in the non-rejection area.

It appears that RPs for Welch's t-test and WMW test are approximately higher than RP for the t-test for the small sample size in the non-rejection area. In the rejection area, RP for t-test is the highest followed by RP for Welch's t-test, then RP for the WMW test. This is because Welch's t-test is designed to handle unequal variances, while the WMW test is non-parametric and does not assume any specific distribution, thus they are robust to violations of assumptions such as Normality and equal variance which are more pronounced in small sample sizes. As the sample size increases, the RPs for the WMW test are smaller than those for Welch's t-test and t-test in the non-rejection area. However, in the rejection area, all three tests have approximately similar RP values.

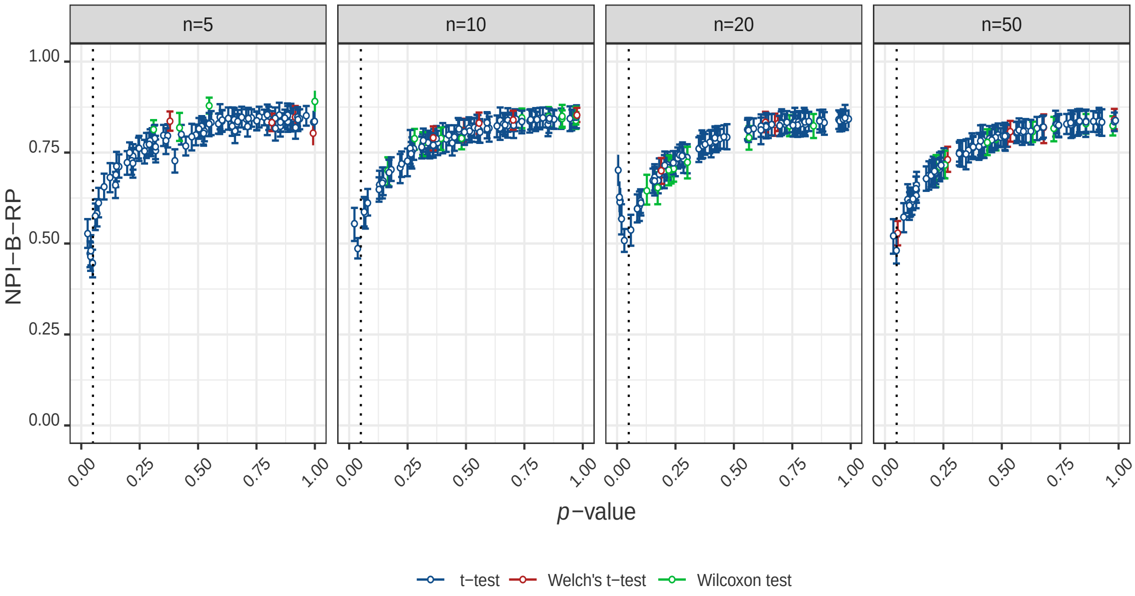


Figure 5.10: The min, mean, and max of RP values for the three-stage tests against p-values for the location test stage, samples are drawn from $N(0,1)$, *Case C*.

Figure 5.16 presents simulation results for the RP of location tests without the preliminary tests. The simulations were conducted with original samples drawn from non-Normal distributions with the same means and variances $LN(0,1)$ (where $\mu_X = \mu_Y \approx 1.648$ and $\sigma_X^2 = \sigma_Y^2 \approx 4.690$). The majority of the original samples are located in the non-rejection area since the means of the two populations are equal. The results show a general pattern of RP. There exists noticeable variability in RP for location tests when dealing with small sample sizes. However, as the sample size increases, this variability decreases. It seems RPs for Welch's t-test and the WMW test are higher than RP for the t-test in the small sample size in the non-rejection area. As the sample size increases, RPs for the WMW test are smaller than RPs for Welch's t-test and t-tests in the non-rejection area.

Figure 5.17 presents simulation results for the RP of the location tests without the preliminary tests. The simulations were conducted with original samples drawn from non-Normal distributions with different means and variances $LN(0,1)$ and $LN(1,0.5)$ (where $\mu_X \approx 1.648, \mu_Y \approx 3.490$ and $\sigma_X^2 \approx 4.690, \sigma_Y^2 \approx 7.899$). Since the original samples are drawn from non-Normal distributions with different means, with increasing the sample

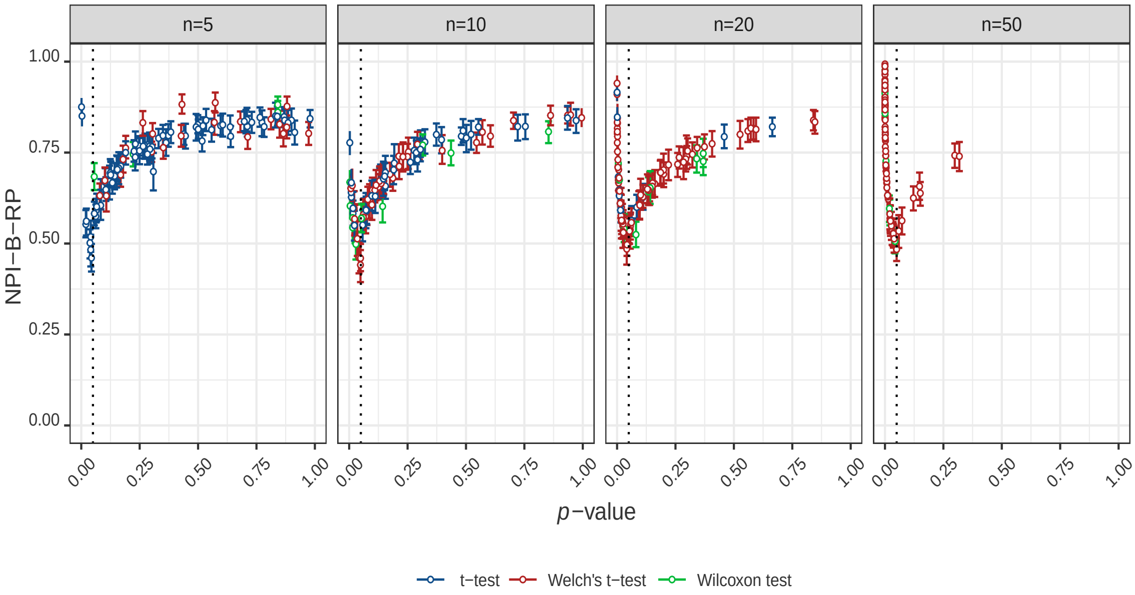


Figure 5.11: The min, mean, and max of RP for the three-stage procedure against p-values for location test, samples are drawn from $N(0,1)$ and $N(1,2^2)$, *Case C*.

size, most original samples shift to the rejection area with high RP values. Whereas in the non-rejection area, p-values become close to the threshold with low RP values. Increasing the size of the sample enhances the RP value for the WMW test, unlike parametric tests which do not exhibit substantial enhancements in RP values within the rejection area. Additionally, there are more original samples located in the non-rejection area for parametric tests compared to those for the WMW test.

There exists noticeable variability in RP for the location tests when dealing with small sample sizes. As the sample size increases, this variability decreases. It seems RPs for the WMW test are the highest in the non-rejection area followed by Welch's t-test. While in the rejection area, as the sample size increases, it is clear that the RPs of the WMW test perform better than the RPs of the parametric test, as they get closer to one in the rejection area than those for the parametric test. This is because the WMW test is more powerful when dealing with non-Normal distribution than the parametric test.

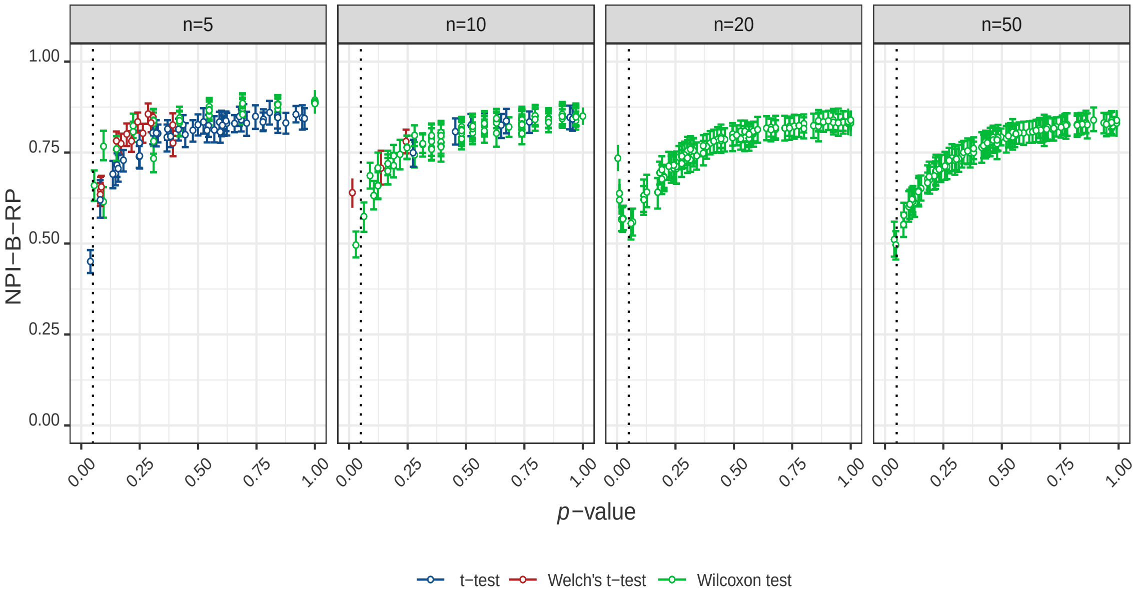


Figure 5.12: The min, mean, and max of RP for the three-stage procedure against p-values for location test, samples are drawn from $LN(0,1)$, *Case C*.

Sample size	t-test		Welch's t-test		WMW test	
	RP	Power	RP	Power	RP	Power
$n = 5$	0.600	0.158	0.600	0.138	0.570	0.111
$n = 10$	0.581	0.273	0.570	0.251	0.562	0.241
$n = 20$	0.682	0.494	0.680	0.486	0.665	0.469
$n = 50$	0.794	0.877	0.792	0.873	0.800	0.850

Table 5.3: The relationship between RP and power for location tests without the preliminary test of Normality, samples are drawn from $N(0,1)$ and $N(1,2^2)$.

The relationship between RP and the estimated power for the location tests without preliminary tests

Tables 5.3 and 5.4 present the relationship between the estimated power for location tests without the preliminary tests with the overall mean of RP values in the rejection area, under the alternative hypothesis for location tests. Generally, it is clear that as the power increases, the RP increases. Moreover, as the sample size increases, the power and RP for location tests increase. When distributions are Normal there is no substantial difference in RP and power among location tests. When distributions are non-Normal and the sample

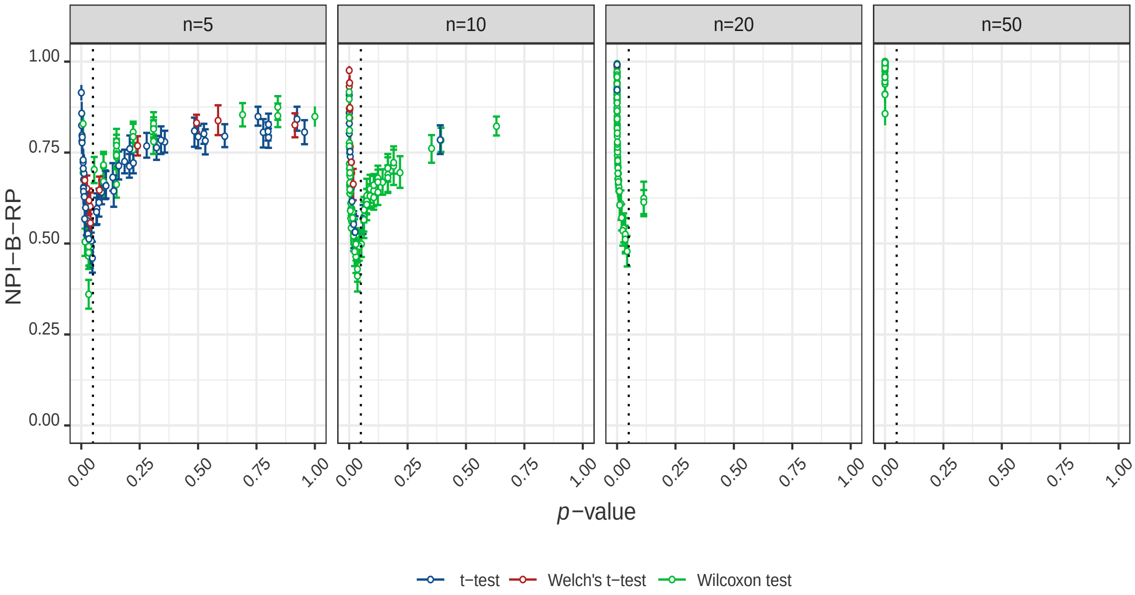


Figure 5.13: The min, mean, and max of RP for the three-stage against p-values for location test, samples are drawn from $LN(0,1)$ and $LN(1,0.5)$, *Case C*.

Sample size	t-test		Welch's t-test		WMW test	
	RP	Power	RP	Power	RP	Power
$n = 5$	0.649	0.369	0.618	0.344	0.588	0.344
$n = 10$	0.711	0.453	0.701	0.549	0.693	0.694
$n = 20$	0.780	0.710	0.778	0.722	0.823	0.961
$n = 50$	0.849	0.920	0.849	0.913	0.987	1.000

Table 5.4: The relationship between RP and power for location tests without the preliminary test of Normality, samples are drawn from $LN(0,1)$ and $LN(1,0.5)$.

size is large, RP and power for the WMW test are slightly better than parametric tests.

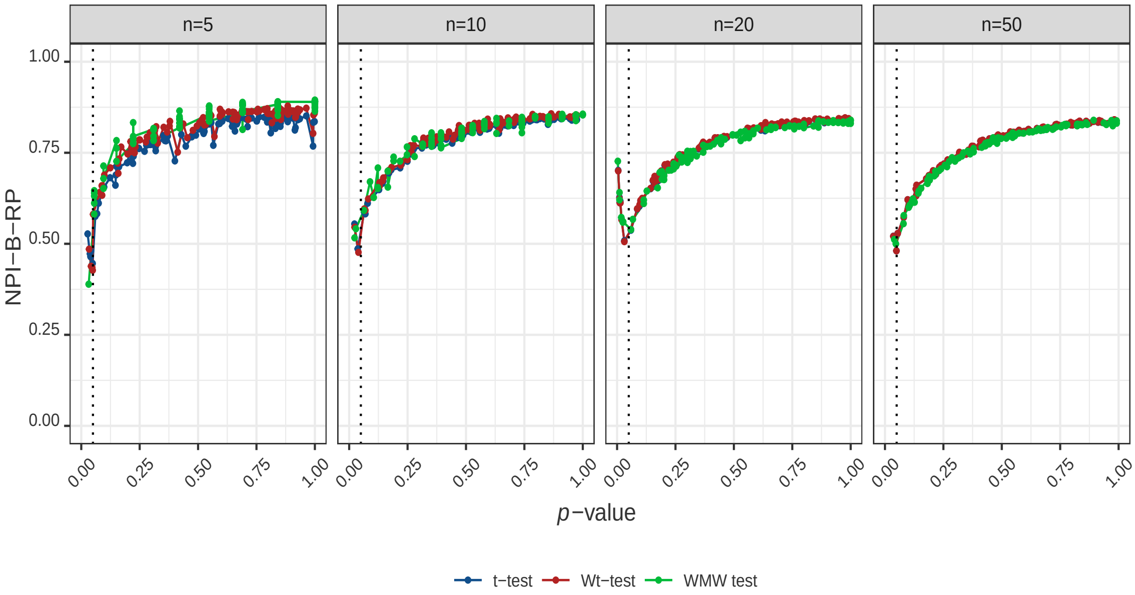


Figure 5.14: The means of RP values for the location tests against their p-values, when original samples are sampled from $N(0, 1)$.

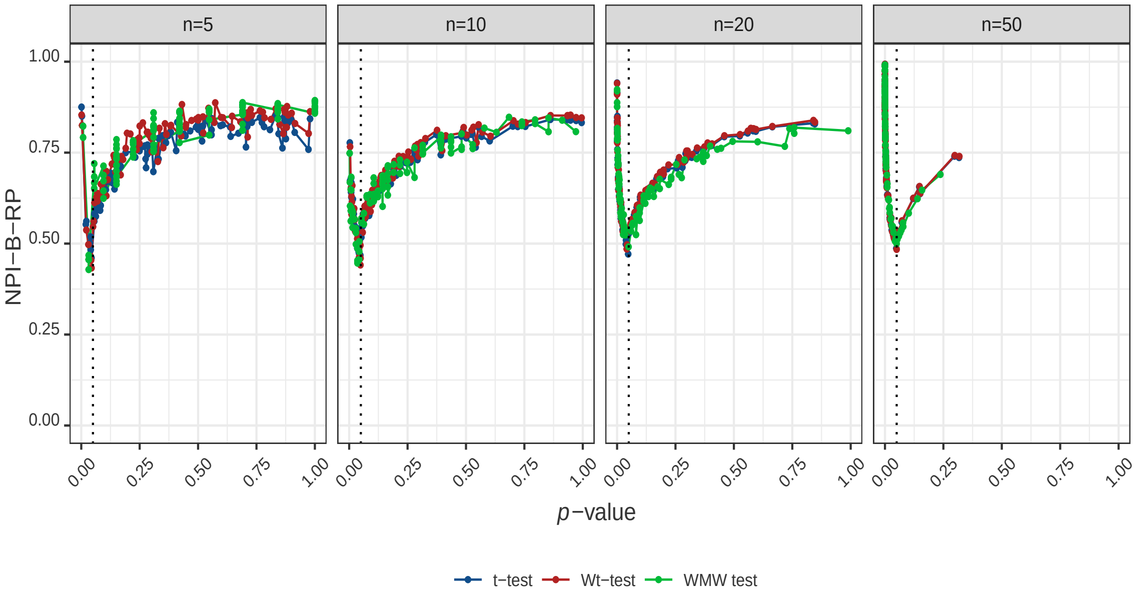


Figure 5.15: The means of RP values for the location tests against their p-values, when original samples are sampled from $N(0, 1)$ and $N(1, 2^2)$.

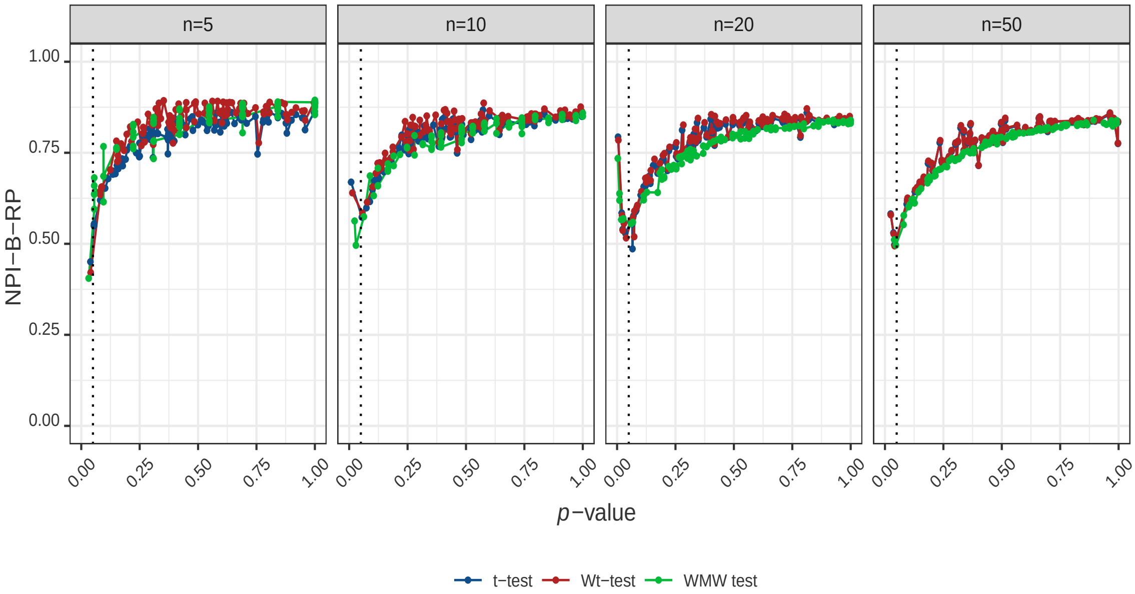


Figure 5.16: The means of RP values for the location tests against their p-values, when original samples are sampled from $LN(0, 1)$.

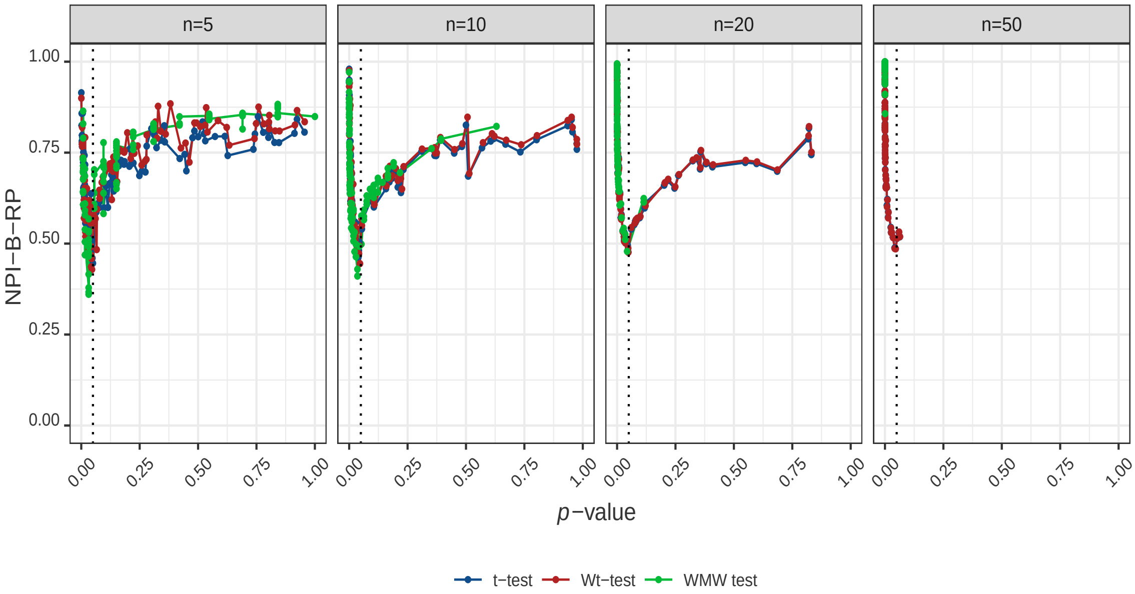


Figure 5.17: The means of RP values for the location tests against their p-values, when original samples are sampled from $LN(0, 1)$ and $LN(1, 0.5)$.

5.5 The impact of the preliminary tests on the reproducibility of location tests

This section includes a comparative analysis between the reproducibility obtained through the three-stage procedure and those obtained through location tests without preliminary tests. The comparison is performed as discussed in Section 4.4. By comparing the RP of the three-stage procedure with the RP for location tests without the preliminary tests, we can assess the impact of the preliminary tests of Normality and equality of variances on the RP of the two-sample location tests.

These comparisons are displayed visually in plots, where the comparison includes comparing the light blue circle representing the RP values of the t-test with the preliminary tests with the dark blue circle indicating the RP values of the t-test without the preliminary tests, also comparing the light green square which represents the RP values of the WMW test with preliminary tests with the dark green square depicting the RP values of the WMW test without preliminary tests, and comparing the light red triangle representing the RP values of Welch's t-test with preliminary tests with the dark red triangle represents the RP values of Welch's t-test without preliminary tests.

5.5.1 The impact of the preliminary tests of Normality and equality of variances on RP of location tests for *Case A*

The results of comparing full RP for the three stages (*Case A*) with the product of the overall mean of the individual RP of location tests and the preliminary tests are presented. The effect of applying preliminary tests for Normality and equality of variances on RP for location tests for *Case A* are shown in Figures 5.18 - 5.21.

We observed that the product of RP for the WMW test and RP of the Normality test tends to be slightly higher than RP for the WMW test with preliminary tests in most scenarios. The difference between them decreases as the sample size increases, especially when the distributions are non-Normal. This is because the two-stage procedure for the two-sample results in increased error rates from both the Normality test and the WMW test. If either of the NPI-B samples fails incorrectly to pass the Normality test this may

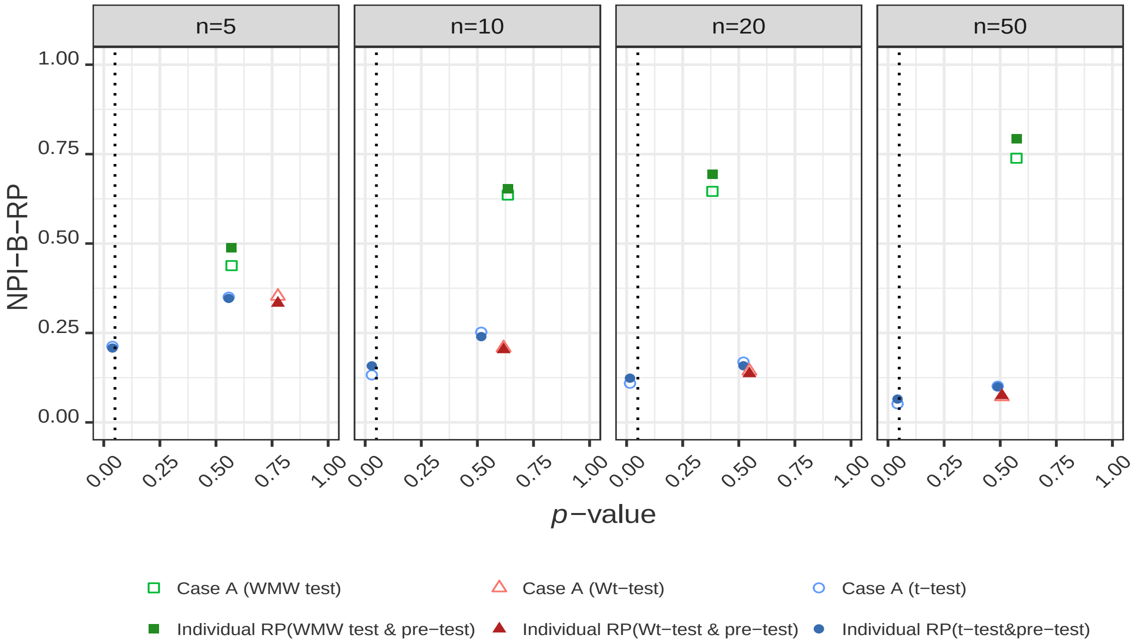


Figure 5.18: Comparing the means of RP values for *Case A* of the three-stage procedure, when both samples are sampled from $N(0, 1)$.

lead to the unnecessary application of the WMW test. Error aggravation can result in a lower overall RP compared to the product of individual RPs, as each test is considered independently. The power of the Normality test increases with the sample size increase, leading to a decreased error rate and then improved RP for the two-stage procedure. There is a slight difference between RP for the t-test with preliminary tests and the product of RP for the three individual tests (Normality and equality of variances test and t-test). Similarly, for RP for Welch's t-test. There is no specific pattern to this difference. Generally, the difference between RP for location tests with preliminary tests and the product of RP of individual location tests and the preliminary tests is considered small. This suggests that the reproducibility of the location tests, conditional on the outcomes of the Normality test and equality of variances test, is not substantially better than that of the location test alone.

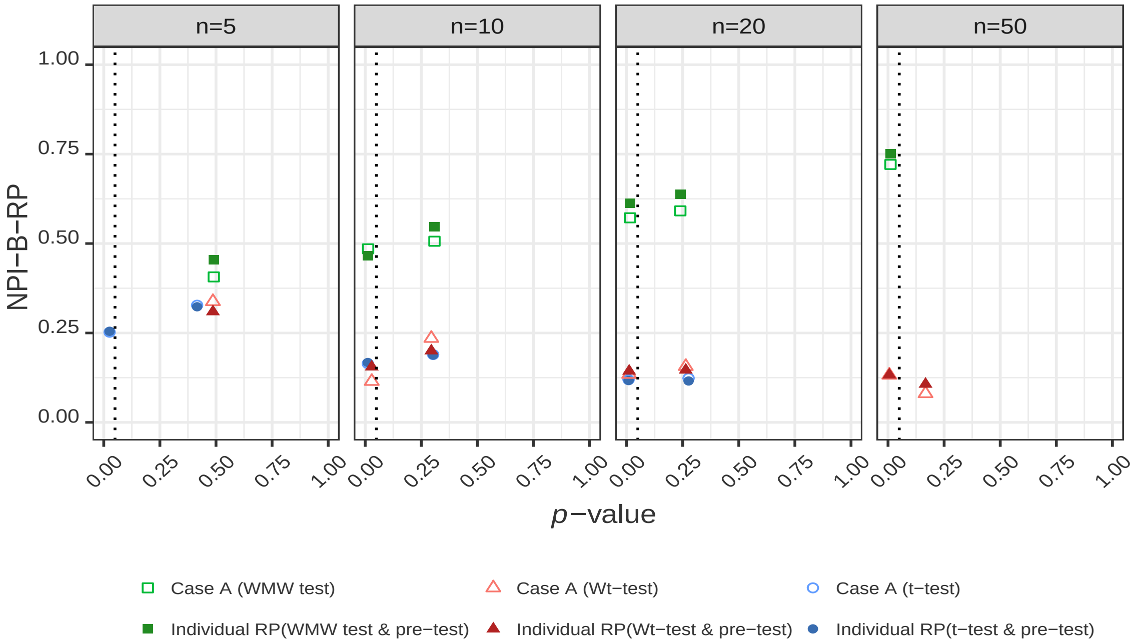


Figure 5.19: Comparing the means of RP values for *Case A* of the 3-stage procedure, when samples are sampled from $N(0,1)$ and $N(1,2^2)$.

5.5.2 The impact of the preliminary tests of Normality and equality of variances on RP of location tests for *Case B*

The results of comparing the reproducibility of the outcome of location tests regardless of which test is used (*Case B*) with the reproducibility of location tests without the preliminary tests are presented. The impact of preliminary tests on RP of location tests for *Case B* is very small as shown in Figures 5.22 - 5.25.

Under the null hypothesis for the location tests in the non-rejection area, the original samples that pass the preliminary tests and do the t-test have a slightly higher overall mean of RP values than RP for the t-test without preliminary tests. As the sample size increases, RP for the t-test with preliminary tests decreases until becomes slightly lower than that without preliminary tests. While the original samples that pass the Normality test and fail to pass F-test and do Welch's t-test have a slightly lower overall mean of RP values than RP for Welch's t-test without preliminary tests. As the sample size increases, the RP for Welch's t-test with preliminary test increases until becomes slightly higher than that without preliminary tests. Whereas the original samples that fail to pass the

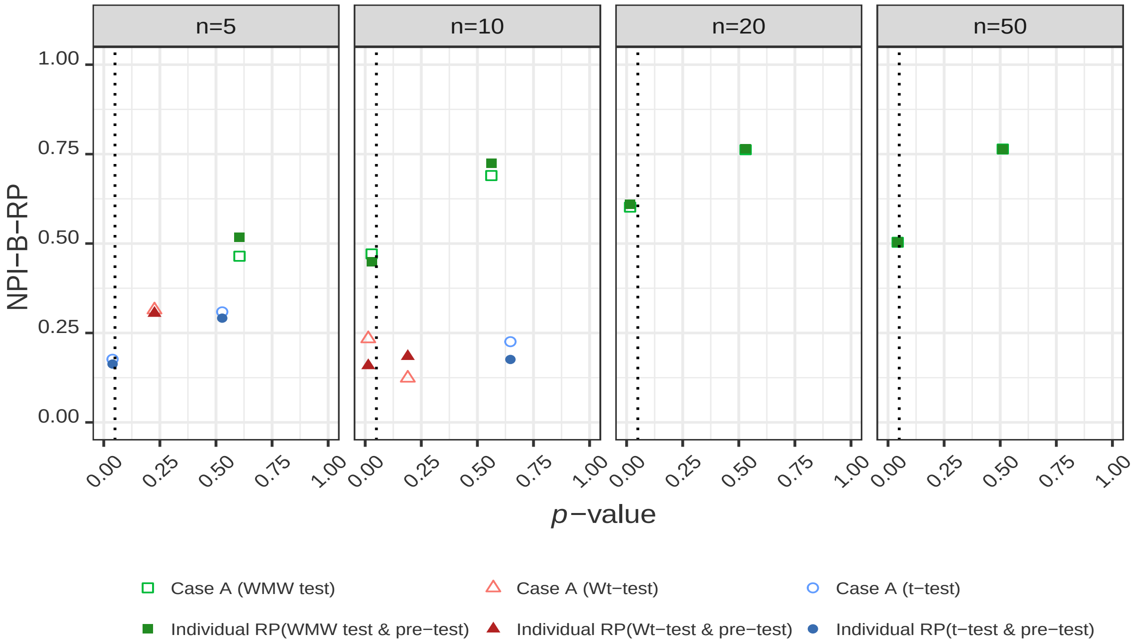


Figure 5.20: Comparing the means of RP values for *Case A* of 3-stage procedure, when samples are sampled from $LN(0,1)$.

Normality tests and do the WMW test have a slightly lower overall mean of RP values than RP for the WMW test without Normality tests. As the sample size increases, RP for the WMW test with preliminary tests convergence to that without preliminary tests until they approximately become similar.

Under the alternative hypothesis for the location tests in the rejection area, RP for the t-test with preliminary tests is smaller than RP for the t-test without preliminary tests. As the sample size increases, RP for the t-test with and without preliminary test becomes approximately similar. RP for Welch's t-test with the preliminary tests is slightly higher than that without the preliminary tests. As the sample size increases, the RP for Welch's t-test with the preliminary test decreases till becomes slightly smaller than that without the preliminary t-test. While RP for the WMW test with and without preliminary tests are approximately similar.

When the distributions are Normal and have different means and variances, RP for Welch's t-test without preliminary tests is slightly higher than RP for Welch's t-test with preliminary tests in the non-rejection area as shown in Figure 5.23.

This small difference between RP for location tests with and without preliminary tests

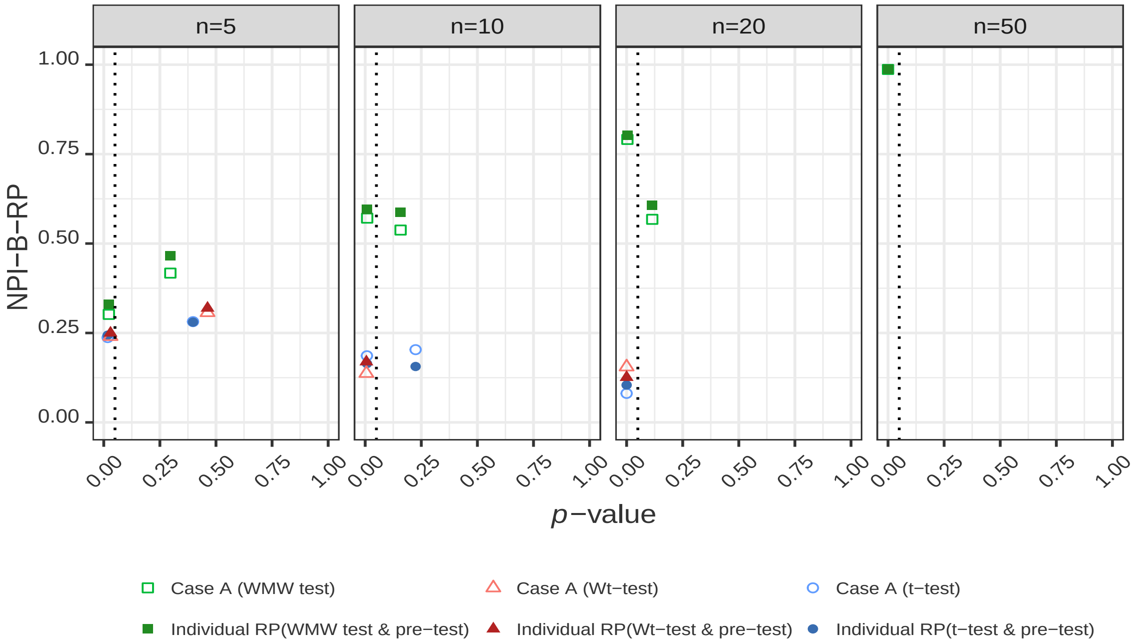


Figure 5.21: Comparing the means of RP values for *Case A* of the 3-stage procedure, when samples are sampled from $LN(0,1)$ and $LN(1,0.5^2)$.

can be traced back to the number of NPI-B samples that pass or fail the preliminary tests and the power of each test on small and large sample sizes. Which enhances or decreases RP for the other location test. Where RP for *Case B* depends on NPI-B samples that have the same outcome as the original sample whether perform t-test Welch's t-test or WMW test. Generally, the difference between the reproducibility of location tests with and without the preliminary test is very small. Thus, the preliminary tests for the Normality and equality of variances do not substantially enhance the reproducibility of the location test outcomes.

5.5.3 The impact of the preliminary tests of Normality and equality of variances on RP of location tests for *Case C*

The comparison results for RP for *Case C* and RP for location tests without preliminary tests are presented. We aimed to determine whether filtering the original samples based on the Normality test and equality of variances results and applying the location test to all NPI-B samples without further preliminary testing could improve the RP of the

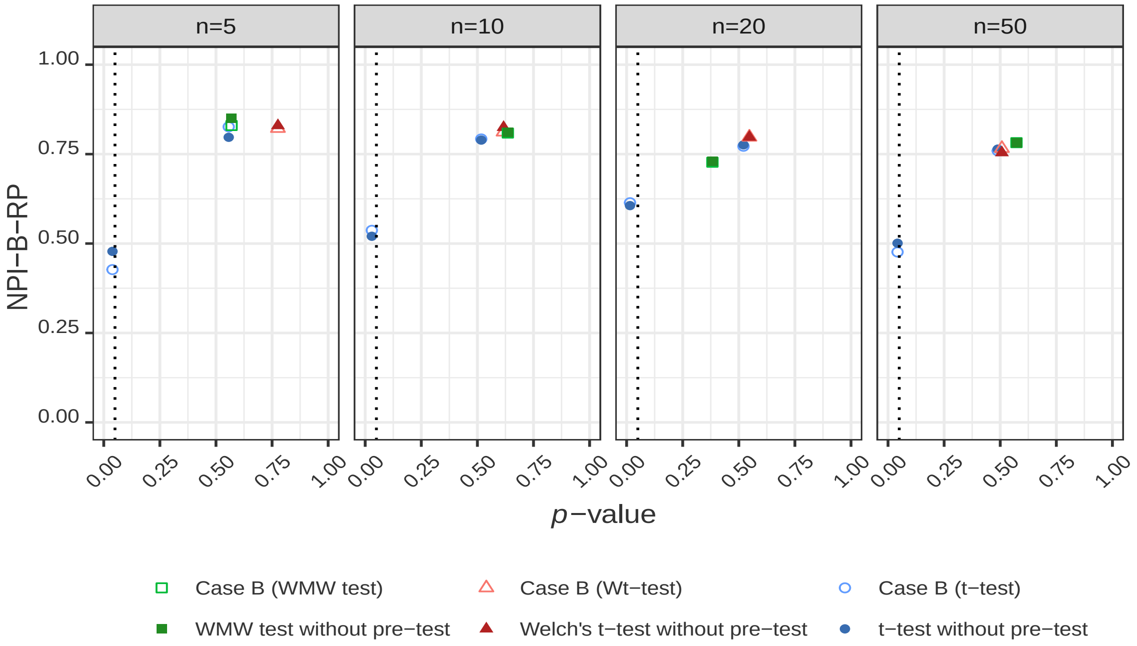


Figure 5.22: Comparing the means of RP values for *Case B* of the three-stage procedure, when both samples are sampled from $N(0, 1)$.

original location test results. By comparing the overall mean of RP values for the t-test after filtering the original samples based on passing the Normality test and equality of variances test with the overall mean of RP values for the t-test without the preliminary tests for all original samples. Likewise, for Welch's t-test, we compare the overall mean of RP values for the Welch's t-test after filtering the original samples based on passing the Normality test and failure to pass the equality of variances test with the overall mean of RP values for the t-test without the preliminary tests for all original samples. In addition, comparing the overall mean of RP values for the Wilcoxon test after filtering the original samples based on failure to pass the Normality test with the overall mean of RP values for the Wilcoxon test without the preliminary tests for all original samples.

Figures 5.26 - 5.29 show the results of the comparison. The RP values for the parametric tests are less affected by the preliminary tests if the distributions are Normal. If the variances are equal, the RP for the t-test is less affected by preliminary tests than the RP for Welch's t-test, as shown in Figure 5.26. However, if the variances are different, the RP for Welch's t-test is less affected than the RP for the t-test, as observed in Figure 5.27. The RP for the WMW test is less affected by the preliminary tests if the distributions are

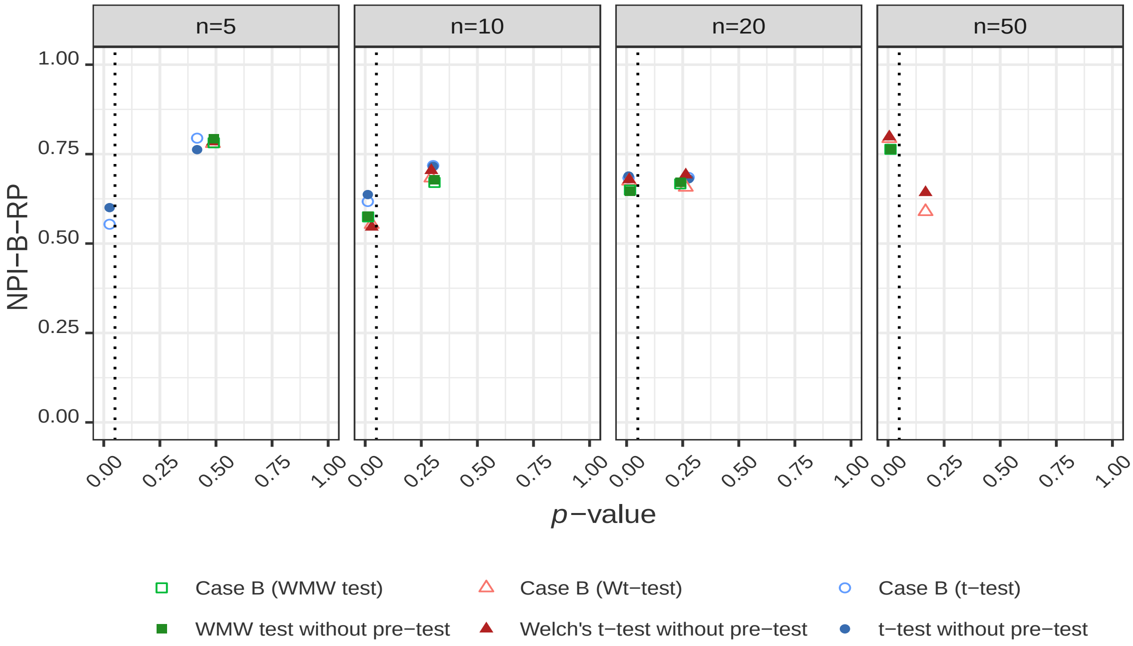


Figure 5.23: Comparing the means of RP values for *Case B* of the three-stage procedure, when samples are sampled from $N(0, 1)$ and $N(1, 2^2)$.

non-Normal, especially in the large sample sizes, as shown in Figures 5.28 and 5.29. This means that filtering original samples that meet the assumptions of the preliminary tests results leads to RP for location tests close to RP for location tests without preliminary tests.

In Figure 5.29, RP for parametric tests with the preliminary tests is higher than those without preliminary tests for sample sizes 10 and 20. This is because the number of original samples that perform these parametric tests after being filtered according to the preliminary tests results is much smaller than the total number of original samples that perform parametric tests without preliminary tests.

Generally, the impact of filtering the original samples based on the preliminary test results on RP of location tests is small and does not improve the RP substantially more than the RP for location tests without the preliminary tests.

Similar results were also achieved when the simulation was conducted for a bimodal distribution from a mixture of normal distributions. The detailed results are provided in Appendix D.1.

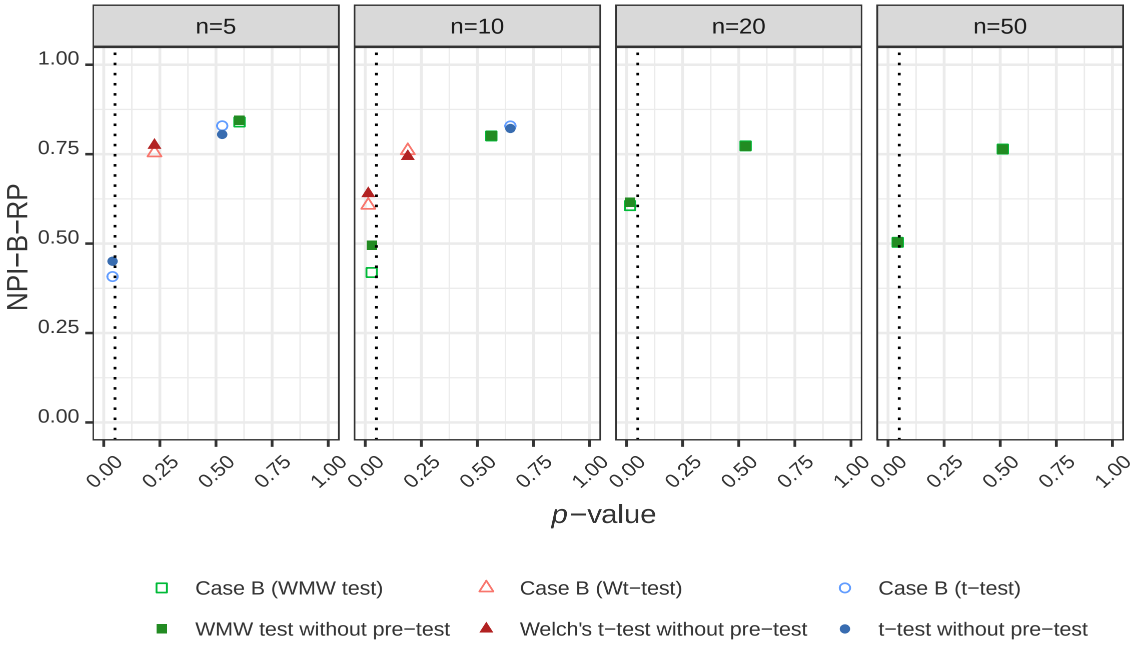


Figure 5.24: Comparing the means of RP values for *Case B* of the three-stage procedure, when samples are sampled from $LN(0, 1)$.

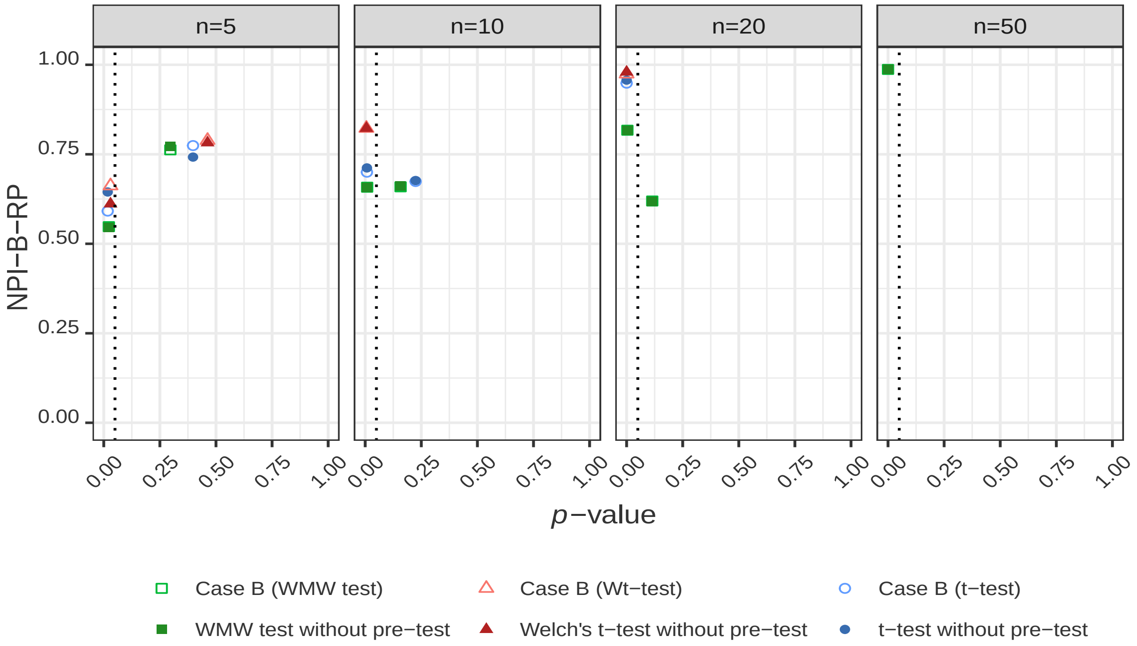


Figure 5.25: Comparing the means of RP values for *Case B* of the three-stage procedure, when samples are sampled from $LN(0, 1)$ and $LN(1, 0.5)$.

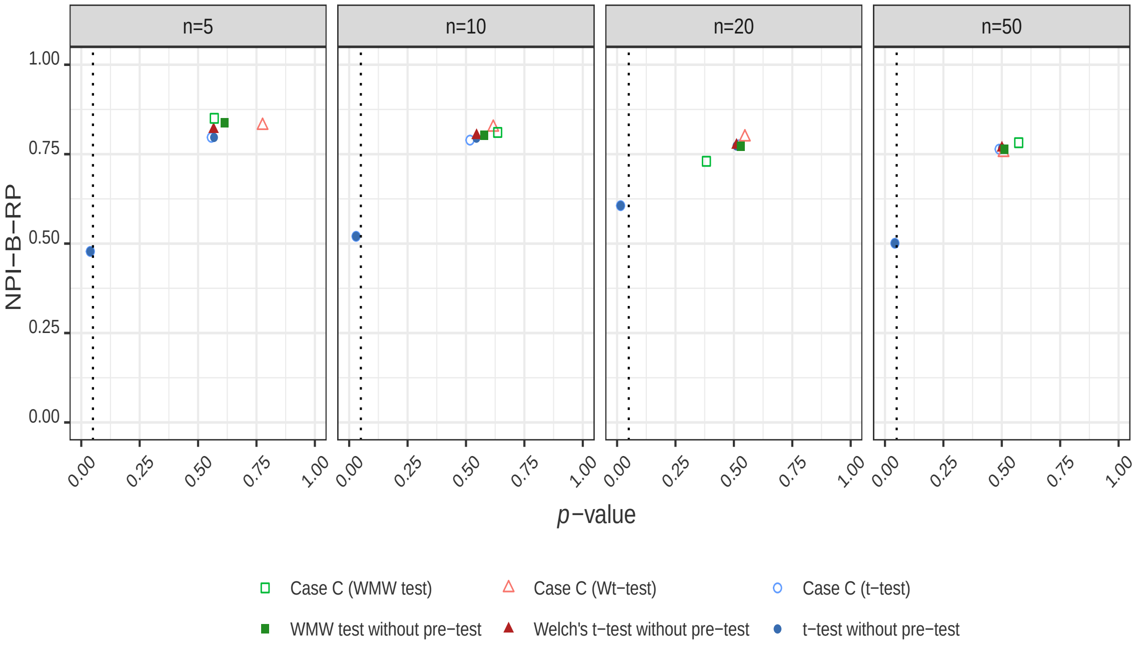


Figure 5.26: Comparing the means of RP values for *Case C* of the three-stage procedure, when both samples are sampled from $N(0,1)$.

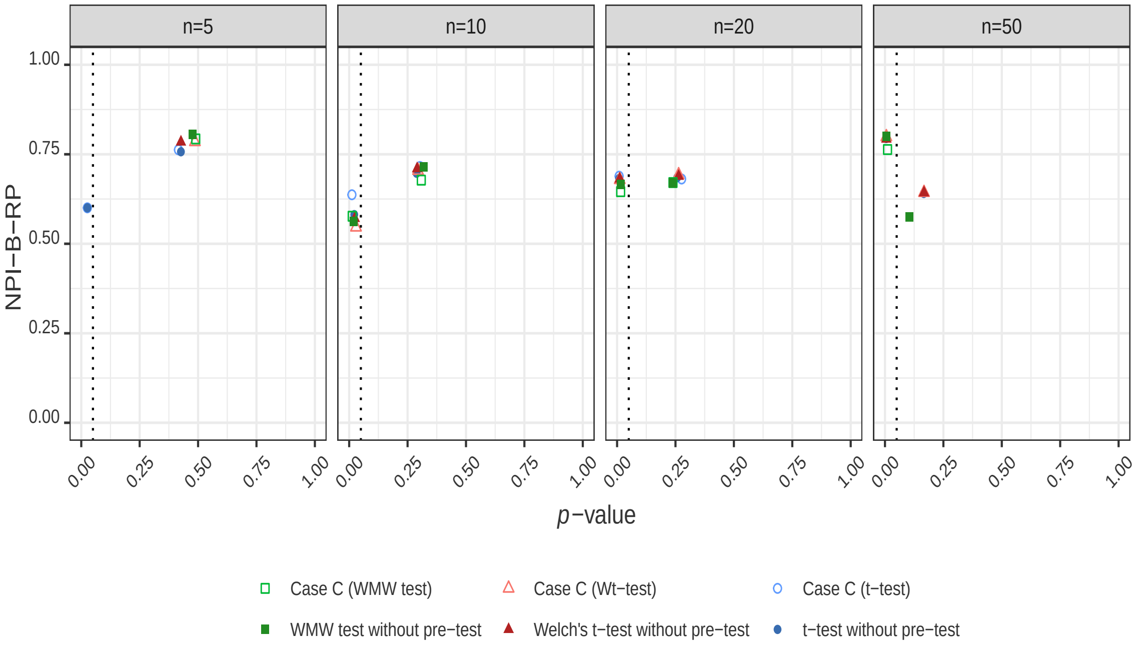


Figure 5.27: Comparing the means of RP values for *Case C* of the three-stage procedure, when samples are sampled from $N(0,1)$ and $N(1,2^2)$.

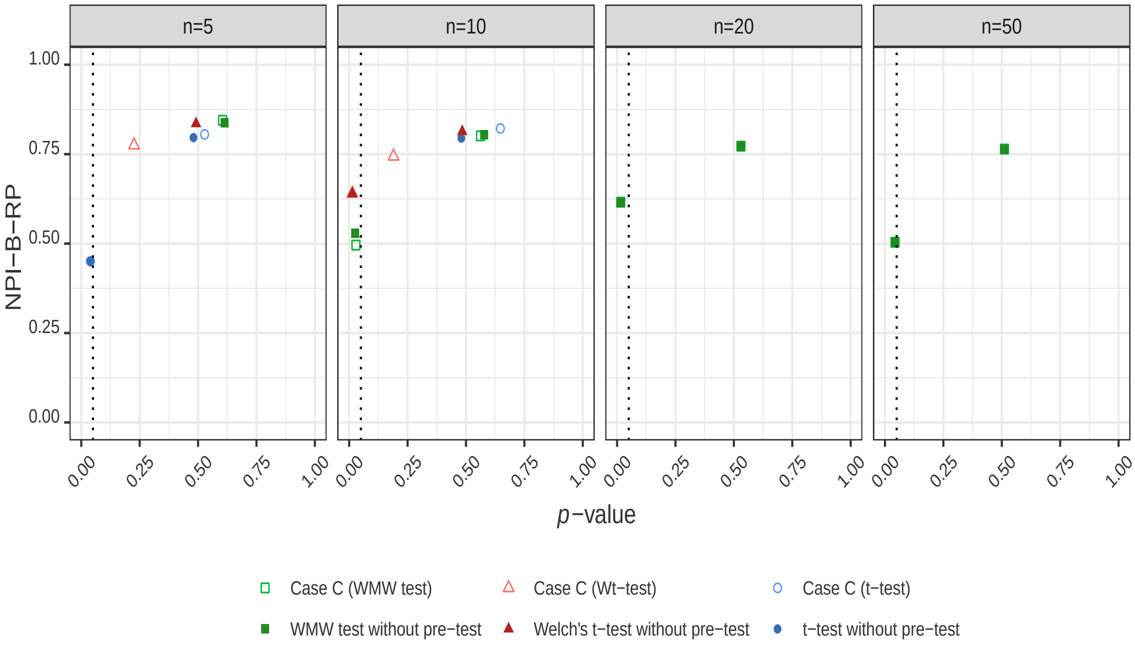


Figure 5.28: Comparing the means of RP values for *Case C* of the three-stage procedure, when samples are sampled from $LN(0, 1)$.

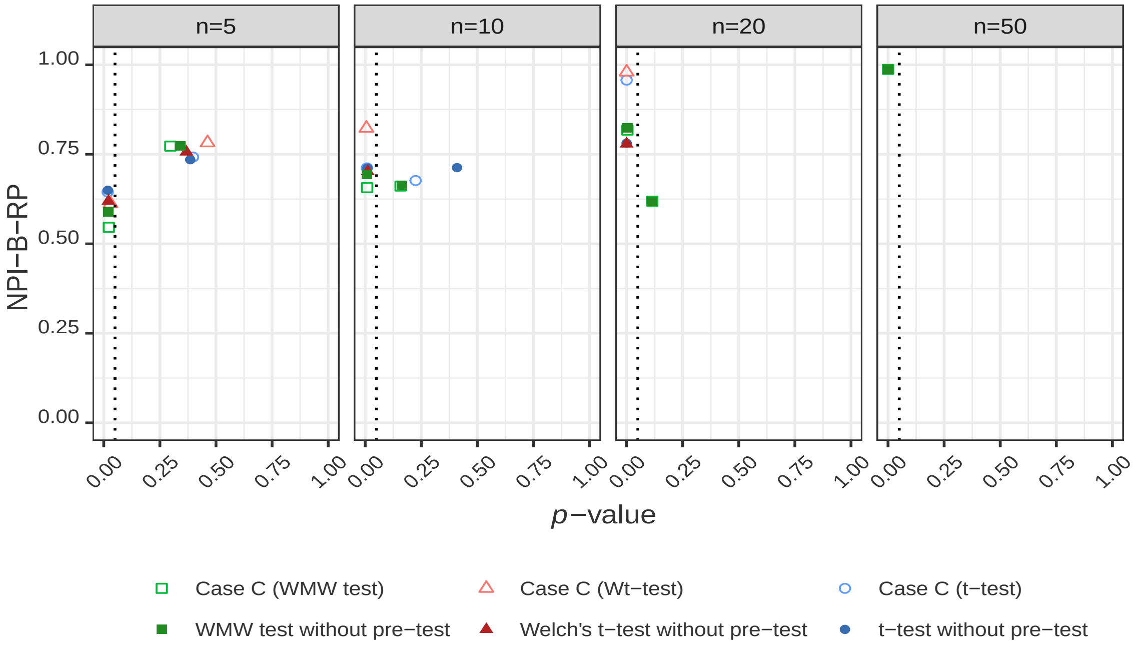


Figure 5.29: Comparing the means of RP values for *Case C* of the three-stage procedure, when samples are sampled from $LN(0, 1)$ and $LN(1, 0.5)$.

5.6 Conclusion

In this chapter, we investigate the reproducibility probability (RP) of the three-stage procedures for two groups. The three-stage procedure involves applying the Normality test in the first stage. Subsequently, if the null hypothesis for this Normality test is rejected for at least one group, then the WMW test is performed. If the null hypothesis for the Normality test is not rejected for both groups, then the equality of variances test is performed. If the null hypothesis for the equality of variances is not rejected, then the t-test is applied in the third stage. Otherwise, Welch's t-test is used. Moreover, the reproducibility of the two-sample location tests without the preliminary tests is also investigated through simulation studies. The goal is to provide valuable insights into the impact of preliminary tests of Normality and equality of variances on RP of the two-sample location tests by comparing RP for location tests with and without preliminary tests.

Three cases of reproducibility probability for the three-stage procedures were examined: *Case A* represents the full RP for the three-stage procedure. *Case B* represents the reproducibility of the same outcome for the location test, no matter which test is used. *Case C* represents reproducibility of the location test conclusion, where for the NPI-B samples the same location test is applied as for the original sample.

The findings show the RP for the three-stage procedures as well as RP for location tests without preliminary tests follow the general pattern for RP: RP values are low when the p-values for location tests are close to the threshold, and when the p-values are far away from the threshold RP values become high. Moreover, RP for location tests has more variability when the sample size is small, as the sample size increases this variability decreases. From the investigation of the full RP for the three stages for *Case A*, RP values for the WMW test tend to increase, as the sample sizes increase. Conversely, the RP values for the parametric tests decrease with increasing sample sizes.

When comparing the RP for the location tests with and without the preliminary tests for Normality and equality of variances for *Cases A, B*, and *C*, we found that the impact of the preliminary tests of Normality and equality of variances on the RP of location tests is generally small. The RP values for the two-sample location tests with preliminary tests do not worsen or improve compared to the RP values for the two-sample location tests without preliminary tests. The impact of preliminary tests on the RP of the location test diminishes substantially and eventually becomes negligible when dealing with large sample sizes and distributions that substantially deviate from Normality.

Moreover, the relationship between the overall mean of RP values in the rejection area and the estimated power for location tests with and without preliminary tests is investigated. RP increases as the estimated power increases, and both increase as the sample size increases.

Chapter 6

Reproducibility of Multiple-Group Location Tests with and without Preliminary Tests

6.1 Introduction

This chapter is an extension of the study of the reproducibility probability (RP) from the nonparametric predictive inference (NPI) perspective for location tests and the impact of applying preliminary tests in their RP. This chapter focuses on investigating the RP for the two-stage procedure for multiple groups, where a preliminary test for Normality is used to choose between one-way analysis of variance (one-way ANOVA) [59, 79, 92] and Kruskal-Wallis tests [42, 68, 89] as multi-group location tests. Additionally, this chapter studies RP for multiple-group location tests without performing the preliminary test for Normality to examine the effect of a preliminary test of Normality on RP of location test by comparing RP of location tests with and without the preliminary test. Furthermore, this chapter conducts a brief investigation into the RP of the three-stage procedure for multiple-group location tests with preliminary tests of Normality and homogeneity of variances. This procedure involves performing a preliminary test for Normality, followed by the Kruskal-Wallis test if the null hypothesis for Normality is rejected for at least one group. Alternatively, if the null hypothesis of the Normality is not rejected, then a preliminary test for equality of variances is conducted. If the null hypothesis for the equality of variances test is rejected, then Welch's ANOVA test is applied; otherwise, the one-way ANOVA test is used.

One-way or one-factor ANOVA is used to determine whether there are any statistically significant differences among the means of three or more independent groups [92]. The one-way ANOVA generalises the t-test to M groups, where $M \geq 2$ [92]. The null hypothesis of one-way ANOVA is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_M$, against the alternative hypothesis H_1 : Not all μ_i are the same $(i = 1, 2, \ldots, M)$ [92]. The test statistic is

$$F = \frac{\text{Mean sum of square between groups}}{\text{Mean sum of square within groups}} = \frac{SS_B/(M-1)}{SS_W/(n_T - M)} \tag{6.1.1}$$

where

$$SS_B = \sum_{i=1}^{M} n_i(\bar{X}_i - \bar{X})^2$$

and

$$SS_W = \sum_{i=1}^{M} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

where n_T is the total number of observations across all groups, M is the number of groups, n_i is the number of observations in the i-th group, X_{ij} is the j-th observation in the i-th group, $\bar{X}_i$ is the mean of the i-th group, and $\bar{X}$ is the overall mean. The ratio F follows F-distribution with $(M-1, n_T - M)$ degree of freedom.

While the Kruskal-Wallis test is a nonparametric alternative test that does not depend on the Normality assumption. Instead, it is employed to compare medians among multiple groups. The statistical value for the Kruskal-Wallis test is typically obtained by computing the H-statistic, which is based on the ranks of the observations [68].

$$H = \frac{12}{n_T(n_T + 1)} \sum \frac{R_i}{n_i} - 3(n_T + 1) \tag{6.1.2}$$

where R_i is the rank sum for the i-th group.

This chapter is structured as follows: Section 6.2 addresses the essential steps to computing the reproducibility of the two-stage procedure for multiple groups. This reproducibility will be investigated via simulations in Section 6.3. This section also presents the reproducibility results for the two-stage procedure obtained via the simulation studies, the reproducibility of the location tests without preliminary tests, and the relationship between RP and the estimated power for the location tests with and without the preliminary test. Then, Section 6.4 focuses on studying the extent to which the preliminary test affects RP of location tests by comparing between RP of the two-stage procedure and RP of location tests without the preliminary test. Lastly, Section 6.5 focuses on summarising the results of the RP of the two-stage procedure and location tests without preliminary tests and the effect of the preliminary test on the RP of location tests.

6.2 Reproducibility for multiple-group location tests with a Normality preliminary test

This section assesses the reproducibility of location tests for multiple groups with a preliminary test for Normality (two-stage procedure). In this two-stage procedure, the first stage involves testing the Normality assumption for each group using a statistical test such as the Shapiro-Wilk test. The second stage involves testing the location difference between the multiple groups using a statistical test such as one-way ANOVA and the Kruskal-Wallis test. If the null hypothesis for the Normality test is rejected for at least one group, then the Kruskal-Wallis test is used, while if all samples pass the Normality test, then the one-way ANOVA test is applied.

Three cases for studying the reproducibility of the two-stage procedure are considered, these are similar to the cases in Section 4.2 of Chapter 4.

Suppose that N is the number of NPI-B samples and N_A^* and N_K^* are two disjoint subsets of $\{1, 2, \ldots, N\}$, and $N_A^* \cap N_K^* = \emptyset$. N_A^* represents indices for NPI-B samples passing the Normality test and the ANOVA test is used, while N_K^* represents indices for those not passing the Normality test and the Kruskal-Wallis test is performed. To estimate NPI-B-RP for the two-stage procedure for multiple groups follow the next steps:

Step 1: Perform a preliminary test for Normality on $M \geq 2$ original samples separately, at significance level α_1.

Step 2: If the null hypothesis H_0^N of the Normality test is rejected for at least one original sample, perform Kruskal-Wallis test and decide about H_0^K with significance level

α_2, set $TS^K = 1$ if H_0^K is rejected or $TS^K = 0$ if H_0^K is not rejected. If all original samples pass the Normality test, perform one-way ANOVA test and decide about H_0^A with significance level α_2, set $TS^A = 1$ if H_0^A is rejected or set $TS^A = 0$ if H_0^A is not rejected.

Step 3: Draw an NPI-B sample N times form each of the original samples, the same size as the original samples.

- For *Cases A* and *B*, perform the Normality test as the preliminary test on the set of N NPI-B samples. Then, ANOVA test or Kruskal-Wallis test is applied based on the preliminary test decision. Record the location test decision each time: $TS_i^A = 1$ if H_0^A in i-iteration is rejected or $TS_i^A = 0$ if H_0^A is not rejected, or record $TS_j^K = 1$ if H_0^K in j-iteration is rejected, or $TS_j^K = 0$ if H_0^K is not rejected, where $i \in N_A^* \subset \{1, 2, \ldots, N\}$ and $j \in N_K^* \subset \{1, 2, \ldots, N\}$.
- For *Case C*, the ANOVA test is performed on the set of N NPI-B samples if the ANOVA test was applied to the original samples. Each time record the test decision $A_s = 1$ if H_0^A is rejected for s-iteration or record $A_s = 0$ if H_0^A is not rejected, where $s = 1, \ldots, N$. The Kruskal-Wallis test is performed on the set of N NPI-B samples if it was performed on original samples. Each time record the test decision $K_s = 1$ if $H_0^{(K)}$ is rejected, or $K_s = 0$ if $H_0^{(K)}$ is not rejected.

Step 4: Compute the RP based on the test decisions of the NPI-B samples.

1. The RP for *Case A*:

   If all original samples passed the Normality test and the one-way ANOVA test was applied to the original samples the RP is:

$$RP_A = \sum_{i \in N_A^*} \mathbb{I}_{\{TS^A = TS_i^A\}} \frac{1}{N}$$

   where $\mathbb{I}_{\{TS^A = TS_i^A\}}$ is an indicator function that takes the value 1 if $TS^A = TS_i^A$ and 0 otherwise.

If one original sample at least did not pass the Normality test and the Kruskal-Wallis test was applied the RP is:

$$RP_K = \sum_{j \in N_K^*} \mathbb{I}_{\{TS^K = TS_j^K\}} \frac{1}{N}$$

2. The RP for *Case B*:

   If all original samples passed the Normality test and the ANOVA test was applied the RP is

$$RP_A = \left(\sum_{i \in N_A^*} \mathbb{I}_{\{TS^A = TS_i^A\}} + \sum_{j \in N_K^*} \mathbb{I}_{\{TS^A = TS_j^K\}} \right) \frac{1}{N}$$

   If one original sample at least did not pass the Normality test and the Kruskal-Wallis test was applied the RP is

$$RP_K = \left(\sum_{i \in N_A^*} \mathbb{I}_{\{TS^K = TS_i^A\}} + \sum_{j \in N_K^*} \mathbb{I}_{\{TS^K = TS_j^K\}} \right) \frac{1}{N}$$

3. The RP for *Case C*:

   If all original samples passed the Normality test and the ANOVA test was applied the RP is

$$RP_A = \sum_{s=1}^{N} \mathbb{I}_{\{TS^A = A_s\}} \frac{1}{N}$$

   If one original sample at least did not pass the Normality test and the Kruskal-Wallis test was applied the RP is

$$RP_K = \sum_{s=1}^{N} \mathbb{I}_{\{TS^K = K_s\}} \frac{1}{N}$$

Step 5: Perform Steps 3 and 4 in total h times, record the outcomes by RP_{A_k}, or RP_{K_k}, where $k = 1, 2, \ldots, h$.

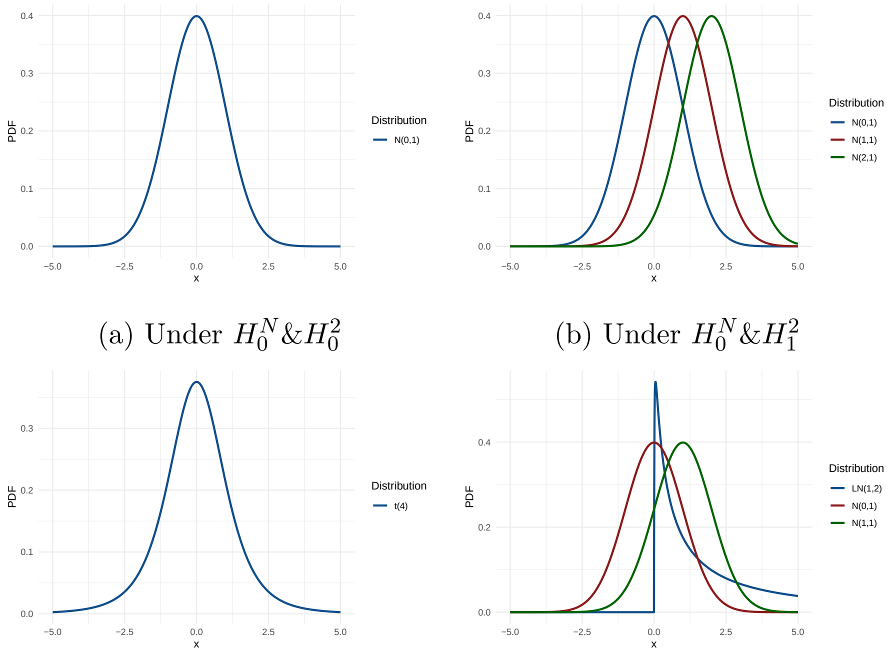
6.3 Simulation studies

We conducted simulations to investigate the reproducibility probability (RP) of the two-stage procedure designed to compare the means or medians of multiple groups. The procedure involves the preliminary stage where the Shapiro-Wilk (SW) test is applied to each sample individually to assess their Normality. If all samples pass the preliminary test, then the one-way ANOVA test is conducted. Otherwise, the Kruskal-Wallis test is employed. To investigate the reproducibility probability for this two-stage procedure, a simulation study is conducted by implementing the algorithm described in Section 6.2 with input parameters $N = 1000$ and $h = 100$. The investigation included two different scenarios involving different numbers of groups: one with $M = 3$ and the other with $M = 5$. Moreover, we conducted simulations to investigate the RP of location tests (one-way ANOVA test and Kruskal-Wallis test) without performing the Normality test, we use the NPI-B-RP Algorithm 1, as presented in Section 1.4.5 of Chapter 1.

The null hypothesis for location tests for multiple groups in the second stage H_0^2, typically states that there are no significant differences in the location parameters θ (e.g. means or medians) among the groups being compared: $H_0^2 : \theta_1 = \ldots = \theta_M$, against the alternative hypothesis H_1^2 : not all θ are equal. H_0^2 represents H_0^A when the ANOVA test is applied in the second stage in which case θ is the means, and H_0^2 represents H_0^K when Kruskal-Wallis test is applied in the second stage in which case θ is the medians.

To conduct the simulations, we consider four scenarios their probability density functions are shown in Figures 6.1 and 6.2: Under H_0^N and H_0^2 (Normality and equal means), for cases where $M = 3$ and $M = 5$, all samples are generated from the standard Normal distribution. Under H_0^N and H_1^2 (Normality and different means), for $M = 3$, we generate data from $N(1,1)$, $N(0,1)$, and $N(2,1)$. For $M = 5$, we generate data from $N(1,1)$, $N(0,1)$, $N(2,1)$, $N(2,2)$, and $N(0,2)$, respectively. Under H_1^N and H_0^2 (non-Normal and equal means), for both $M = 3$ and $M = 5$, all samples are generated from t-distribution with 4 of degree of freedom. Under H_1^N and H_1^2 (non-Normal and different means), in the case of $M = 3$, we select the distributions $N(0,1)$, $N(1,1)$, and $LN(1,2)$, respectively. For $M = 5$, we opt the distributions $N(0,1)$, $N(1,1)$, $LN(1,2)$, $Ca(1,1)$, and $t(4)$.

The number of runs per simulation is 50, with various sample sizes that are the same for each group 5, 15, 30, these numbers were chosen because they present enough information about the pattern of RP values. All the results are presented based on a two-sided test with a significance level of 5% for both stages.



(a) Under $H_0^N \& H_0^2$

(b) Under $H_0^N \& H_1^2$

(c) Under $H_1^N \& H_0^2$

(d) Under $H_1^N \& H_1^2$

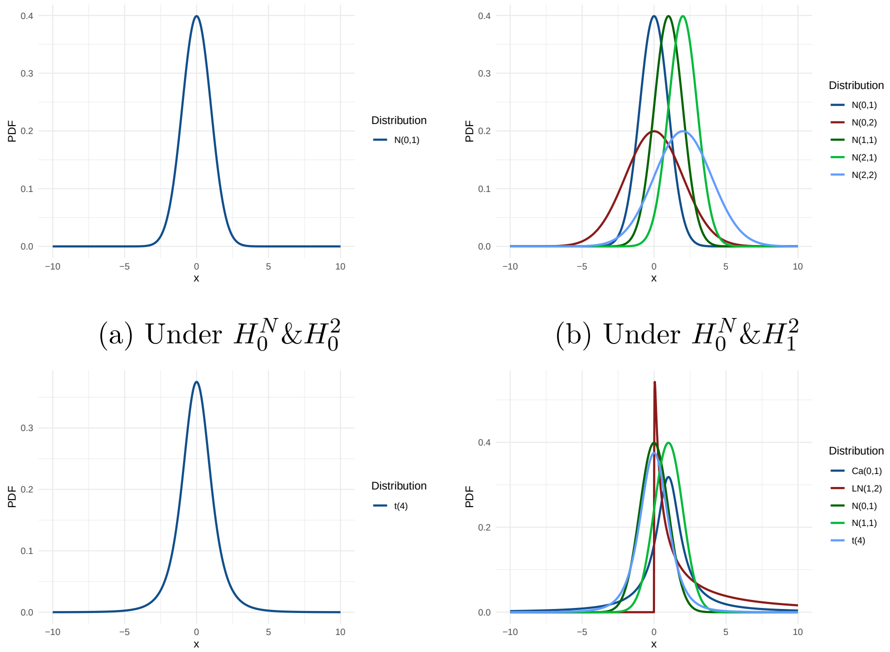
Figure 6.1: PDFs for distributions are considered in the simulation, $M = 3$

6.3.1 Simulation results for the reproducibility for location tests with the preliminary test

In this section, the results of the simulation for NPI-B-RP for the two-stage procedure of multi-group location tests with the preliminary Normality test for *Cases A*, *B* and *C* and for the 3 and 5 groups are presented. The results are represented visually in plots, where the y-axis represents min, mean, and max RP values for the location tests with the preliminary test. In contrast, the x-axis represents the p-values for the location test stage. The blue colour represents RP values for the two-stage procedure, where the original samples pass the Normality test and the ANOVA test is used. The green colour represents RP values where at least one original sample does not pass the Normality test and performs the Kruskal-Wallis test.

The results for *Case A*

Results for full reproducibility for the two-stage procedure are presented. Figures 6.3, 6.5, 6.7 and 6.9 illustrate simulation results for the two-stage procedure for *Case A* involving



(a) Under $H_0^N \& H_0^2$

(b) Under $H_0^N \& H_1^2$

(c) Under $H_1^N \& H_0^2$

(d) Under $H_1^N \& H_1^2$

Figure 6.2: PDFs for distributions that are considered in the simulation, $M = 5$

3 groups. Similarly, Figures 6.4, 6.6, 6.8 and 6.10 show simulation outcomes for 5 groups.

The observed relationship between RP and the p-value for location tests from the simulation study is that RP values tend to increase as the p-value moves further away from the threshold, and when the p-value is close to the threshold, RP values tend to decrease. This trend is particularly pronounced for the nonparametric test (Kurskal-Wallis test) for all sample sizes and the parametric test (ANOVA test) with small sample sizes.

When both the number of groups and sample size are large, the RP for ANOVA tends to have the same value approximately near zero regardless of the location of the p-value whether near the threshold or far, as shown in Figures 6.4, 6.6, and 6.8. Moreover, as the sample size and the number of groups increases, the RP for the ANOVA test decreases, while the RP for the Kruskal-Wallis test increases. This can be attributed to the performance of the Normality test with NPI-B samples of diverse distributions. Specifically, with the larger sample size and the larger number of groups, the chances of all groups of NPI-B samples passing the Normality test decrease. Therefore, the non-parametric test is applied instead of parametric tests, resulting in a reduced RP for

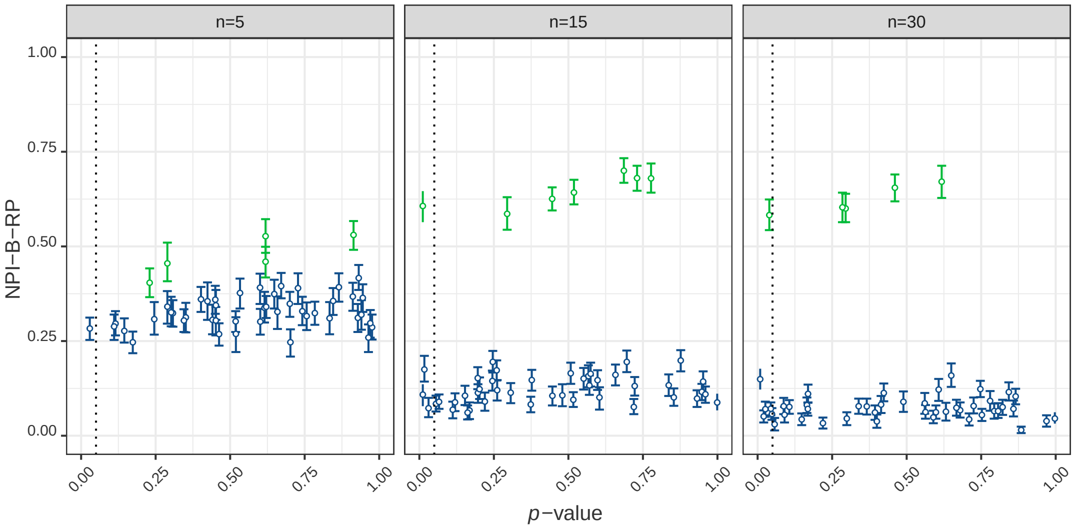


Figure 6.3: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $M = 3$.

ANOVA and an increased RP for the Kruskal-Wallis test.

The results for *Case B*

This section shows results for the reproducibility of the location test's outcome. Figures 6.11, 6.13, 6.15, and 6.17 present the results of simulation studies for the RP of the two-stage procedure for (*Case B*) for 3 groups. While Figures 6.12, 6.14, 6.16, and 6.18 show the results for 5 groups. In this case, the RP whether for the ANOVA test or the Kurskal-Wallis test exhibits a general pattern of RP, there is no difference in the performance of the RP of ANOVA and Kruskal-Wallis tests as it was in *Case A*.

When all original groups are drawn from distributions that satisfy the null hypothesis for both stages (Normality and equal means), as illustrated in 6.11 and 6.12, the majority of original samples pass Normality test and proceed with the ANOVA test in the non-rejection area. In cases where all original groups are drawn from distributions that are Normal but possess different means, as shown in 6.13 and 6.14, the ANOVA test also perform better than the Kruskal-Wallis test, according to the number of original samples, most original samples opt for the ANOVA test after the Normality test.

When all original samples are drawn from distributions that are non-Normal but have the same mean, as illustrated in 6.15 and 6.16, in large sample size the majority of original

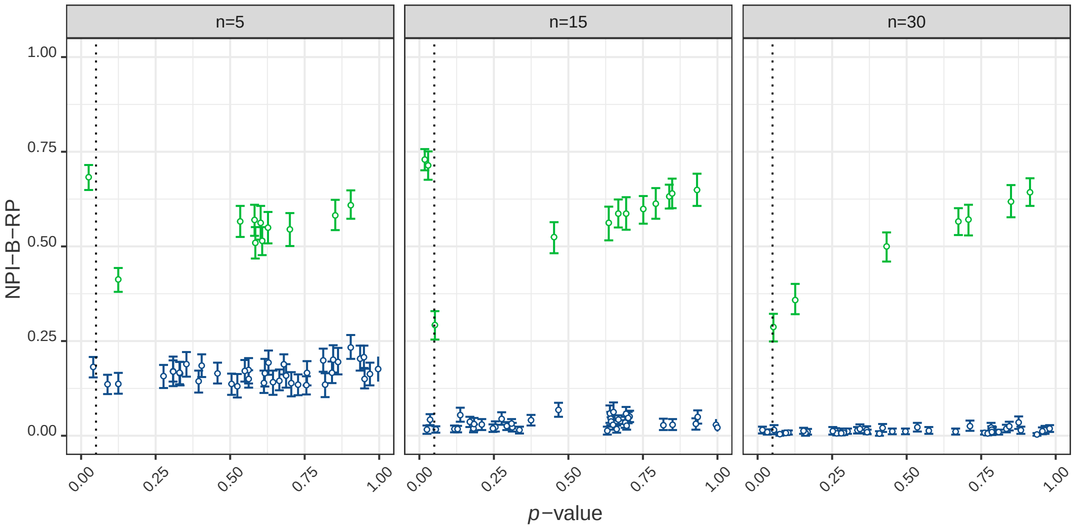


Figure 6.4: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0, 1)$, $M = 5$.

samples perform the Kruskal-Wallis test after applying Normality test in the non-rejection area. However, in the small sample size, it seems that the ANOVA test is applied more than the Kruskal-Wallis test. This is because when dealing with small sample sizes, the Normality test has a lower power to detect deviations from Normality. Additionally, the characteristics of the $t(4)$ distribution, which may not be considered extremely high in kurtosis, can lead to a relatively higher number of sets of original samples passing the Normality test. This, in turn, results in a higher proportion of sets performing the ANOVA test.

When the original groups are drawn from distributions not all are Normal and they possess different means, as shown in 6.17 and 6.18, the Kruskal-Wallis test exhibits performance better than ANOVA test, particularly for large sample sizes. It was observed that the majority of original samples that performed the two-stage procedure seemed to use the Kruskal-Wallis test after applying the preliminary test of Normality.

The RP values of the two-stage testing in *Case B* are affected by the number of groups, with RP values for location tests in the non-rejection area decreasing as the number of groups increases, while those in the rejection area increase. This is traced back to the RP being affected by the power of the test and the effect size (the amount of difference between the groups). As the number of groups increases, a larger effect size is required

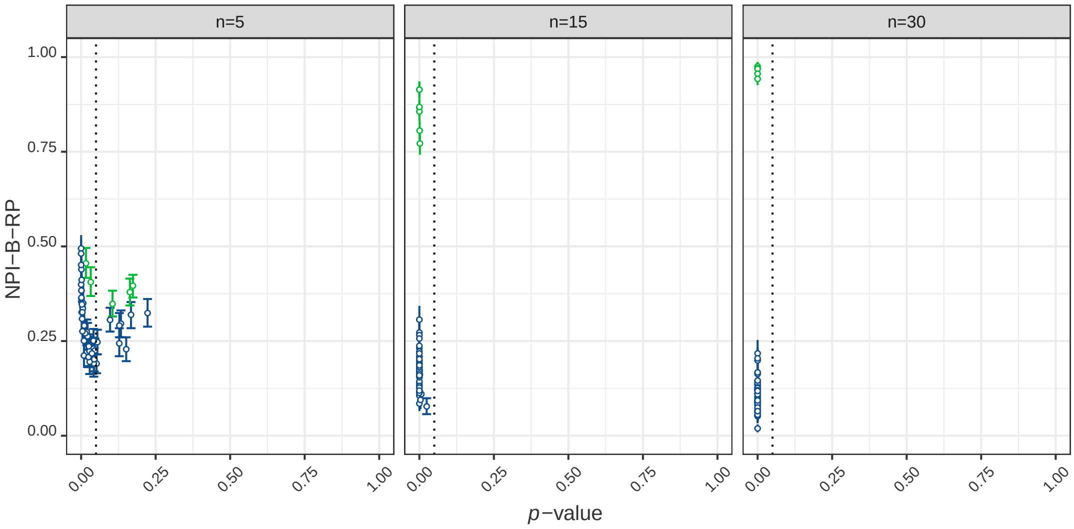


Figure 6.5: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

to achieve statistical significance. On the other hand, the test has higher power when the effect size is large enough to pass the significance level. This indicates that the null hypothesis is more likely to be rejected when a true difference is present. As a result, the RP values are higher in the rejection area. This is because the test consistently identifies a true effect.

The results for *Case C*

Results of simulation studies for RP of the location test conclusion, where for the NPI-B samples the same location test is applied as for the original sample without further the preliminary test are presented. Figures 6.19, 6.21, 6.23, and 6.25 present the results of the simulation study concerning the RP for three groups. Similarly, Figures 6.20, 6.26, 6.24, and 6.22 show the results for five groups. RP values for this case show the general pattern for RP and they exhibit similar patterns observed in *Case B*.

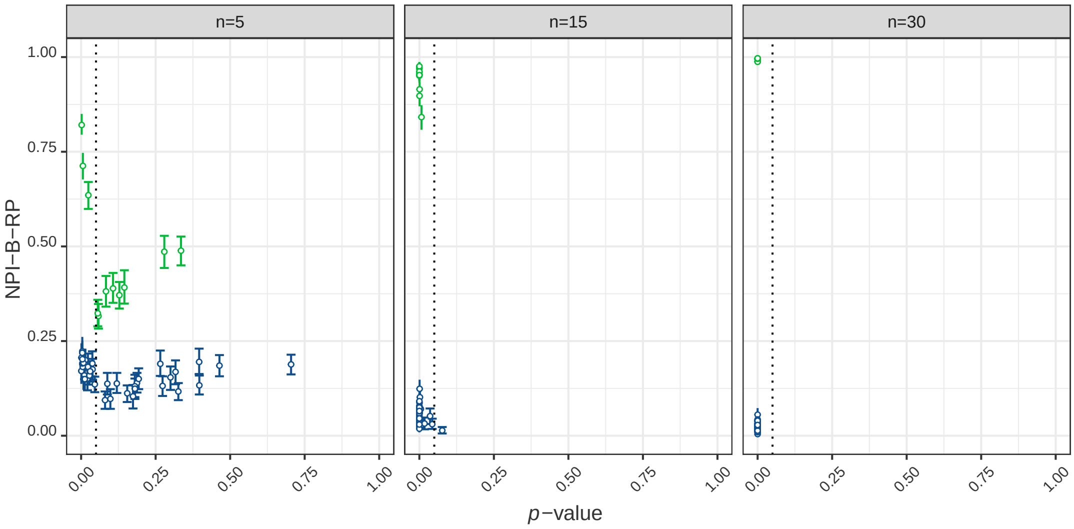


Figure 6.6: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(1,1)$, $N(0,1)$, $N(2,1)$,$N(2,2)$, and $N(0,2)$, $M = 5$.

Sample size	RP	Power
$n = 5$	0.675	0.693
$n = 15$	0.928	0.998
$n = 30$	0.997	1.000

Table 6.1: The relationship between RP and power for the two-stage procedure, samples from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

The relationship between RP and the estimated power for the two-stage procedure

Tables 6.1 and 6.2 present the relationship between the overall mean of RP values in the rejection area and the estimated power for the two-stage procedure that are discussed in Section 4.3.1, for 3 groups, when original samples are from Normal and when not all distributions are Normal, respectively. The relationship between RP and power is positive: RP increases as power increases. Additionally, both RP and power increase as the sample size increases.

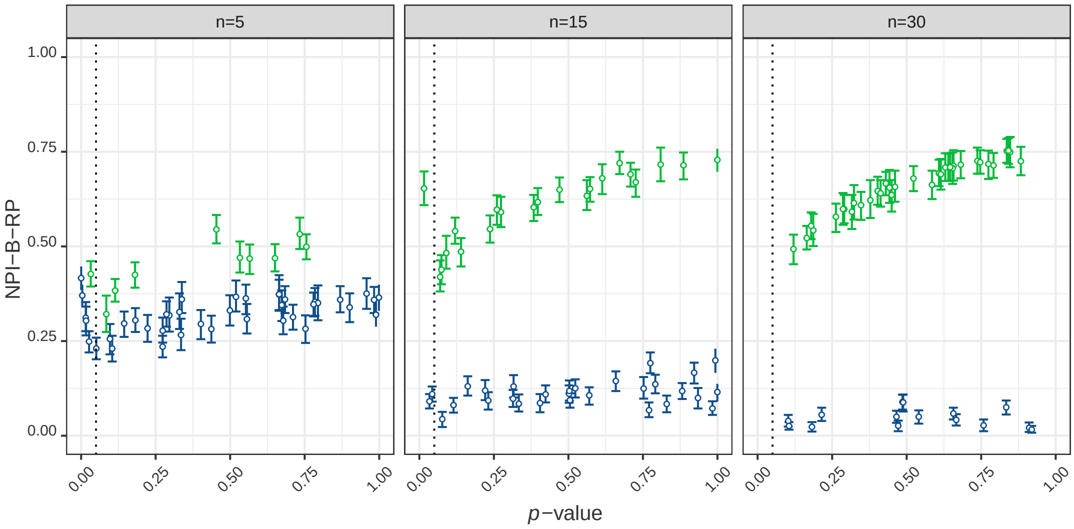


Figure 6.7: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $t(4)$, $M = 3$.

Sample size	RP	Power
$n = 5$	0.695	0.465
$n = 15$	0.887	0.982
$n = 30$	0.992	1.000

Table 6.2: The relationship between RP and power for the two-stage procedure, samples from $N(0, 1)$, $N(1, 1)$, and $LN(1, 2)$, $M = 3$.

6.3.2 Simulation results for the reproducibility of location tests without the preliminary test

The results of the simulation studies for the reproducibility of multi-group location tests, the one-way ANOVA test and the Kruskal-Wallis test, without preliminary tests for 3 and 5 groups, are presented.

Figures 6.27 and 6.28 present the results of RP values for the ANOVA test and Kruskal-Wallis test without the preliminary test of Normality, for 3 groups and 5 groups, respectively. These results are displayed under hypothesis H_0^N and H_0^2 (Normality and equality in means) when all groups are generated from $N(0, 1)$. RP values for both multi-group location tests for 3 and 5 groups show the general pattern for RP: RP is low when the

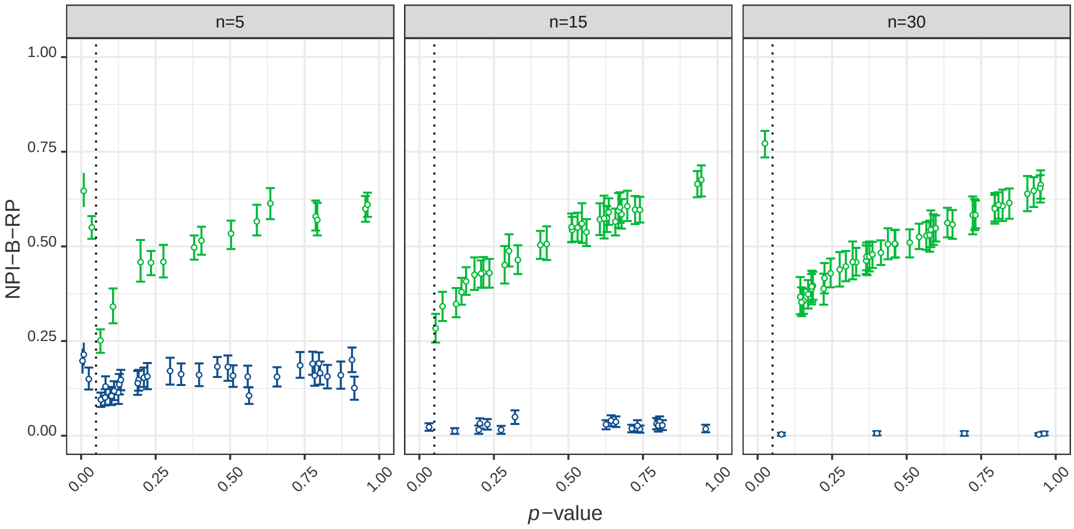


Figure 6.8: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $t(4)$, $M = 5$.

p-value for location tests is close to α, and RP values increase gradually as the p-value moves away from α. In this situation where all distributions are Normal and have the same means, most of the original samples fall into the non-rejection area. When comparing the RP values of the ANOVA test with RP values for the Kruskal-Wallis test, for the small sample size, the Kruskal-Wallis test demonstrates slightly higher RP values compared to those for the ANOVA test. However, with an increase in sample size, the RP values for the ANOVA test become slightly higher than that of the Kruskal-Wallis test. This is because the Kruskal-Wallis test is more robust to small non-Normality NPI-B samples, resulting in slightly higher RP values than the ANOVA test. As the sample size increases, the ANOVA test gains power due to the central limit theorem, leading to higher RP values compared to the Kruskal-Wallis test.

Moreover, RP values for both location tests without the preliminary test exhibit slight variability, particularly for smaller sample sizes. Nevertheless, with a large sample size, this variability tends to decrease. This is due to the low power of tests on small sizes.

As the sample size and the number of groups increase, the RP values for location tests without preliminary tests decrease in the non-rejection area and increase in the rejection area.

Figures 6.29 and 6.30 show the RP values for ANOVA and Kruskal-Wallis tests without

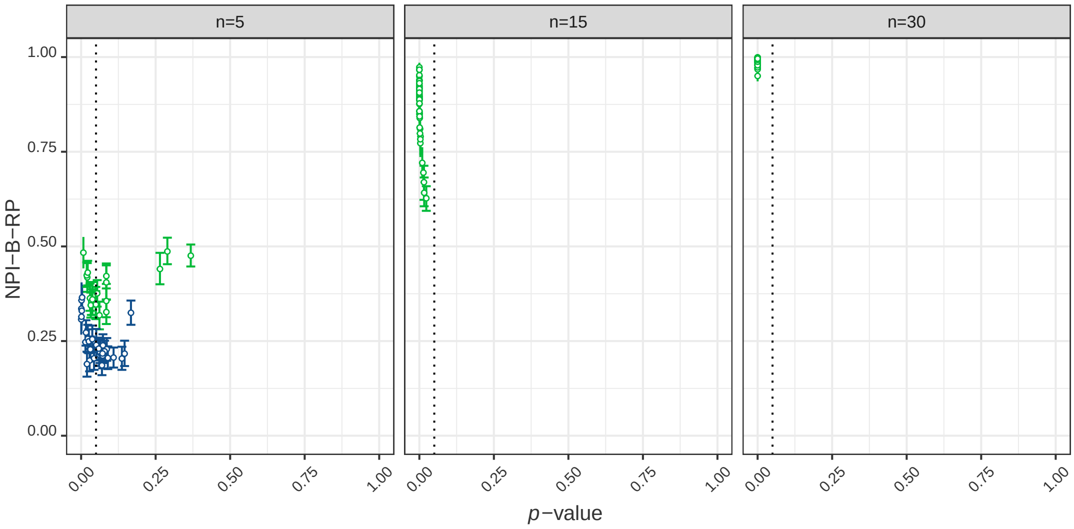


Figure 6.9: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $N(1,1)$, and $LN(1,2)$, $M = 3$.

a preliminary test for 3 and 5 groups, respectively. These results are presented under hypotheses H_0^N and H_1^2 (Normality and differences in means). In this scenario, where all distributions are Normal but have different means, most original samples fall into the rejection area and exhibit high RP values approaching one, indicating significant differences in means among distributions. It appears that when the number of groups is small (3), there is minimal difference in RP between the two location tests across various sample sizes. However, when dealing with a larger number of groups (5), the RP values for the Kruskal-Wallis test are slightly better than those of the ANOVA test. This is because as the number of groups increases, the power of the ANOVA test is affected by any slight deviation from Normality which leads to reduced RP values for the ANOVA test. While the Kruskal-Wallis test maintains higher power because it does not depend on the Normality assumption.

In Figures 6.31 and 6.32, the RP values for the location tests without the preliminary test are presented for 3 and 5 groups, respectively. These results are displayed under H_1^N and H_0^2 (non-Normality and equality in means) the samples are generated from t-distribution with the degree of freedom 4 for both 3 and 5 groups. In this situation where all distributions are not Normal but have the same means, most of the original samples are

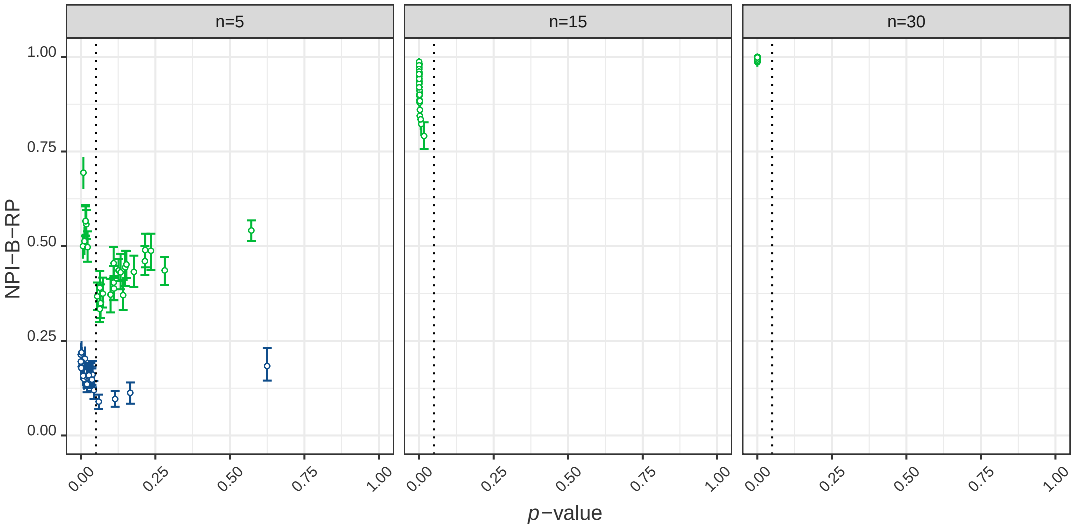


Figure 6.10: RP values for then two-stage procedure *Case A* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $N(1,1)$, $LN(1,2)$, $Ca(1,1)$, and $t(4)$, $M=5$.

located in the non-rejection area, and the RP of the Kruskal-Wallis test is approximately higher than the RP of the ANOVA test for the small sample size. As the sample size increases, RP values for Kruskal-Wallis decrease until they become a little less than RP values for the ANOVA test.

Figures 6.33 and 6.34 display the RP values for the ANOVA test and Kruskal-Wallis test without preliminary test, for 3 and 5 groups, respectively. These results are presented under the hypothesis, H_1^N and H_1^2 (non-Normality and differences in the means). The RP values of the Kruskal-Wallis test are approximately the same as the RP values of the ANOVA test in the rejection area for the small sample size. Increasing the sample size improves the performance of the RP values for the Kruskal-Wallis test in the rejection area, unlike the ANOVA test, which does not show substantial improvements in RP values. Additionally, as the sample size increases, the number of original samples that are located in the rejection area of the ANOVA test is less than that for the Kruskal-Wallis test. This is because the Kruskal-Wallis test has a higher power than the ANOVA test, reflected in the number of original samples located in the rejection area and their RP values.

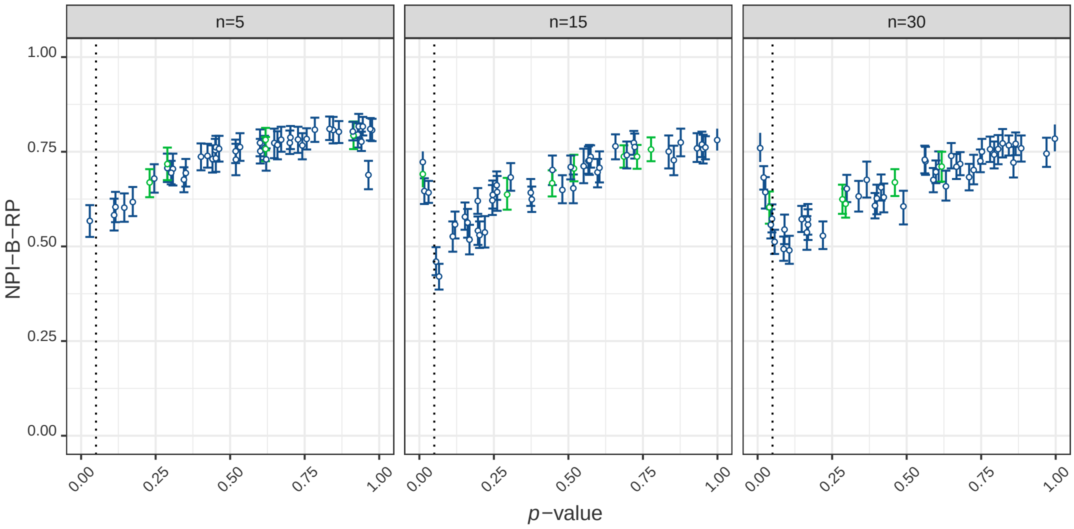


Figure 6.11: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $M = 3$.

Sample size	ANOVA test		Kruskal-Wallis test	
	RP	Power	RP	Power
$n = 5$	0.692	0.701	0.682	0.635
$n = 15$	0.934	0.998	0.927	0.997
$n = 30$	0.997	1.000	0.997	1.000

Table 6.3: The relationship between RP and power for location tests without the Normality test, original samples from $N(1,1)$, $N(0,1)$, and $N(2,1)$, respectively, $M = 3$.

The relationship between RP and the estimated power for the multiple-group location tests

Table 6.3 presents the relationship between the overall mean of RP values of location tests in the rejection area and the estimated power for location tests without preliminary tests under the alternative hypothesis, using samples drawn from Normal distributions for 3 groups. The table illustrates that both RP and power tend to increase with sample sizes, and there is a positive correlation between power and RP.

Furthermore, Table 6.4 explores the relationship between the overall mean of RP values in the rejection area and the power for location tests without preliminary tests, under the alternative hypothesis, using samples not all drawn from Normal distributions

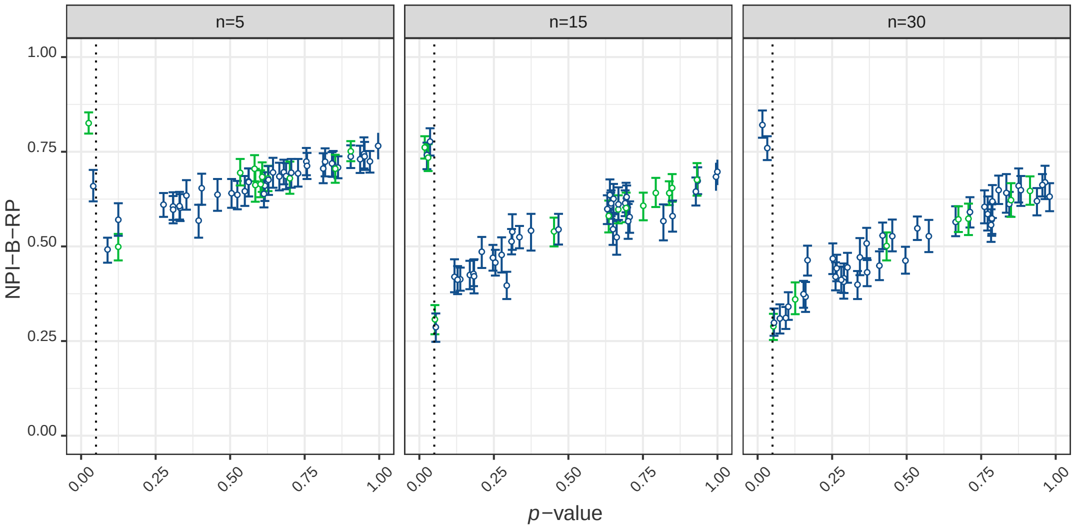


Figure 6.12: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $M = 5$.

Sample size	ANOVA test		Kruskal-Wallis test	
	RP	Power	RP	Power
$n = 5$	0.681	0.229	0.596	0.520
$n = 15$	0.707	0.646	0.889	0.978
$n = 30$	0.793	0.813	0.992	1.000

Table 6.4: The relationship between RP and power for location tests without the Normality test, original sample from $N(0,1)$, $N(1,1)$, and $LN(1,2)$, respectively, $M = 3$.

for 3 groups. It is evident from the table that as the power increases, RP also increases. Moreover, the power and RP of the Kruskal-Wallis test are better than those of the ANOVA test. This observation asserts that the Kruskal-Wallis test's effectiveness in detecting differences across multiple means, particularly in distributions with high kurtosis (e.g., $LN(1,2)$ distribution, which exhibits high kurtosis). Similar findings for power have been reported in previous studies such as [60] and [57].

Similar relationships between the overall mean of RP in the rejection area and the power are found when the number of groups is 5.

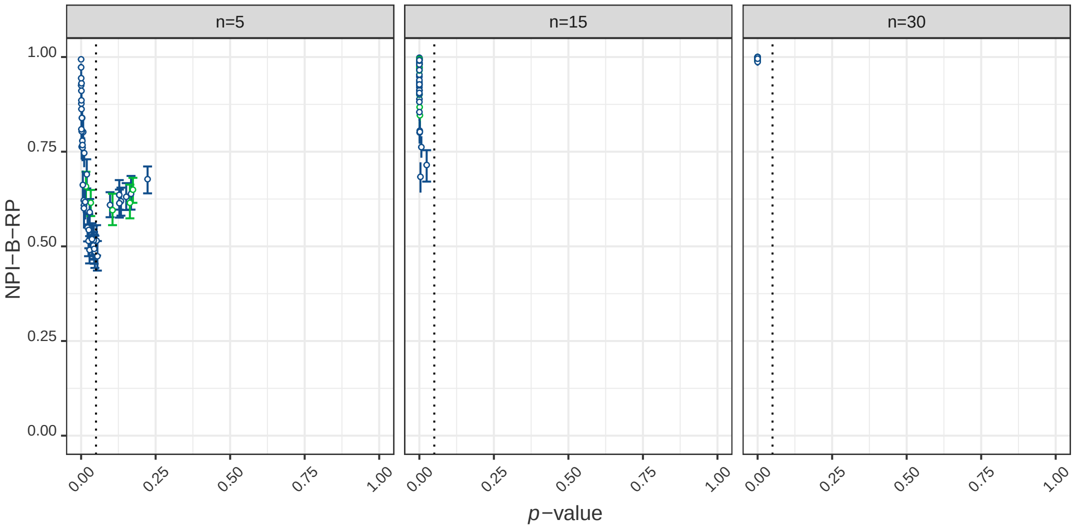


Figure 6.13: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

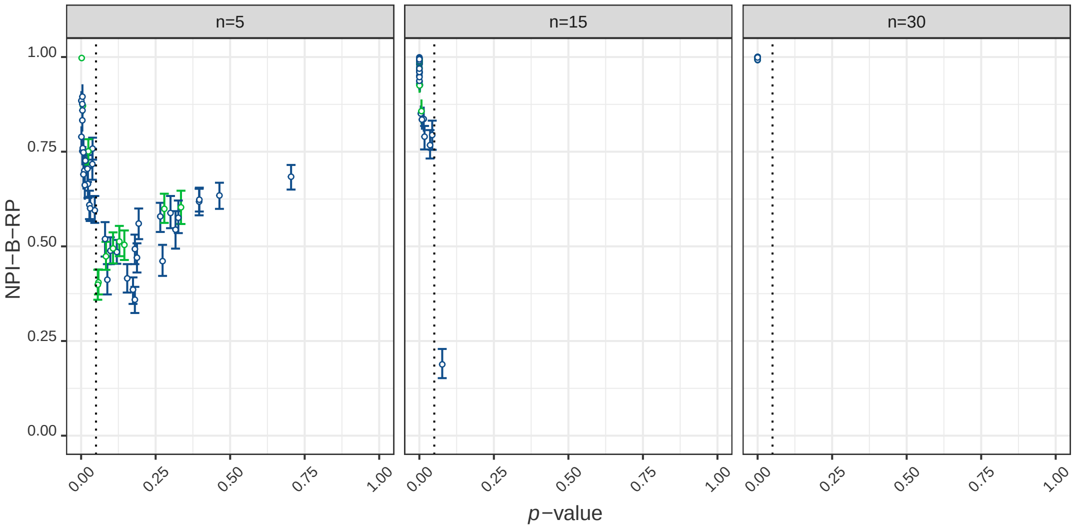


Figure 6.14: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(1,1)$, $N(0,1)$, $N(2,1)$,$N(2,2)$, and $N(0,2)$, $M = 5$.

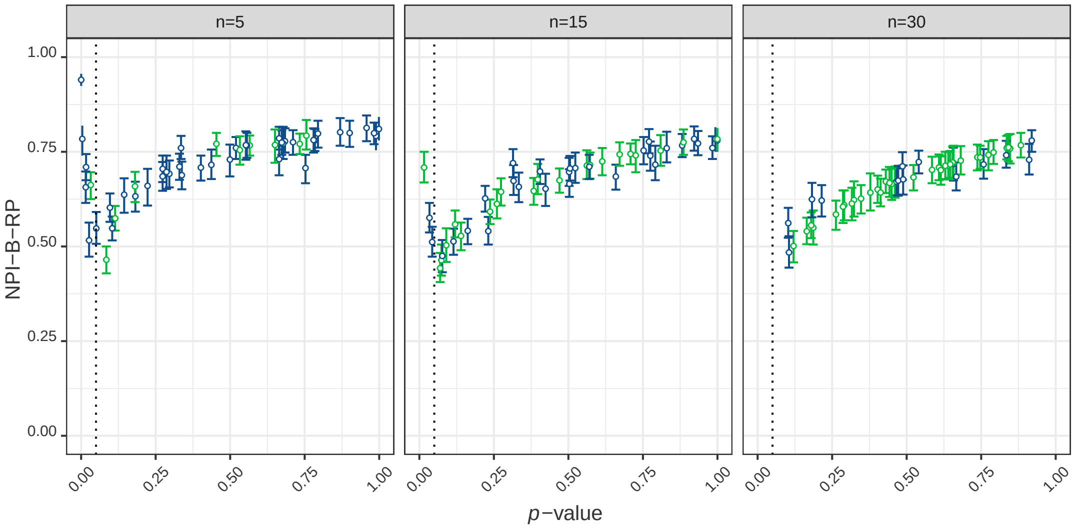


Figure 6.15: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $t(4)$, $M = 3$.

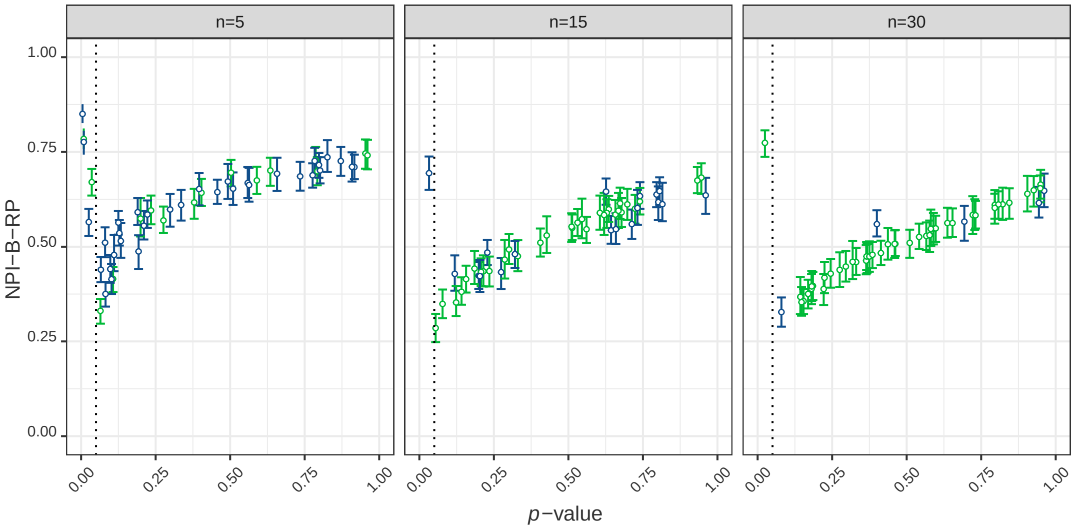


Figure 6.16: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $t(4)$, $M = 5$.

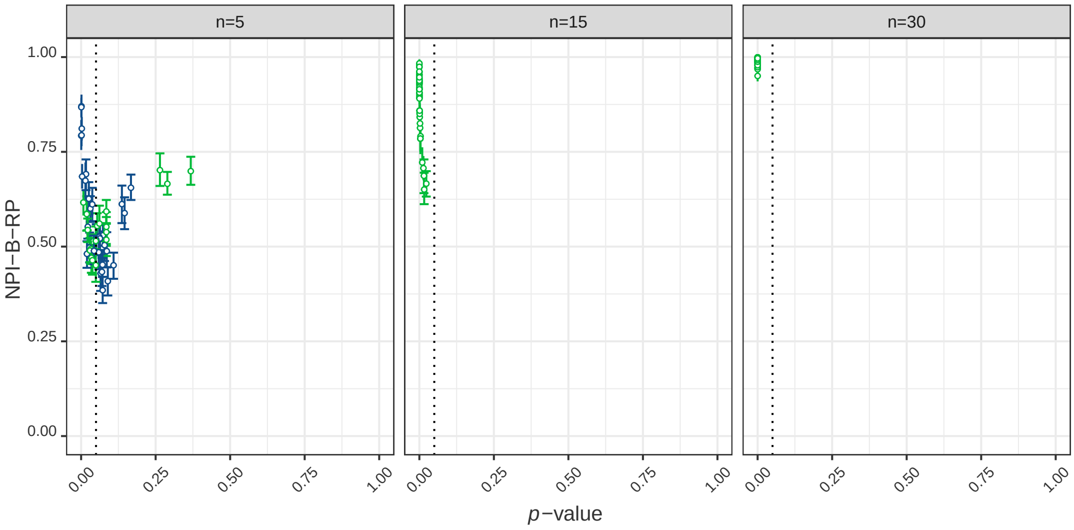


Figure 6.17: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $N(1,1)$, and $LN(1,2)$, $M = 3$.

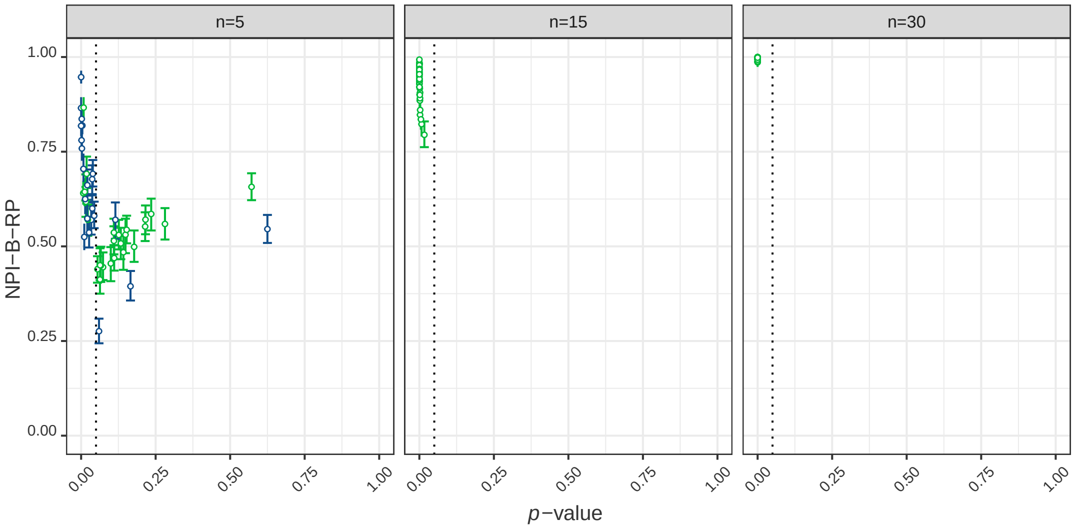


Figure 6.18: RP values for then two-stage procedure *Case B* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0,1)$, $N(1,1)$, $LN(1,2)$, $Ca(1,1)$, and $t(4)$, $M = 5$.

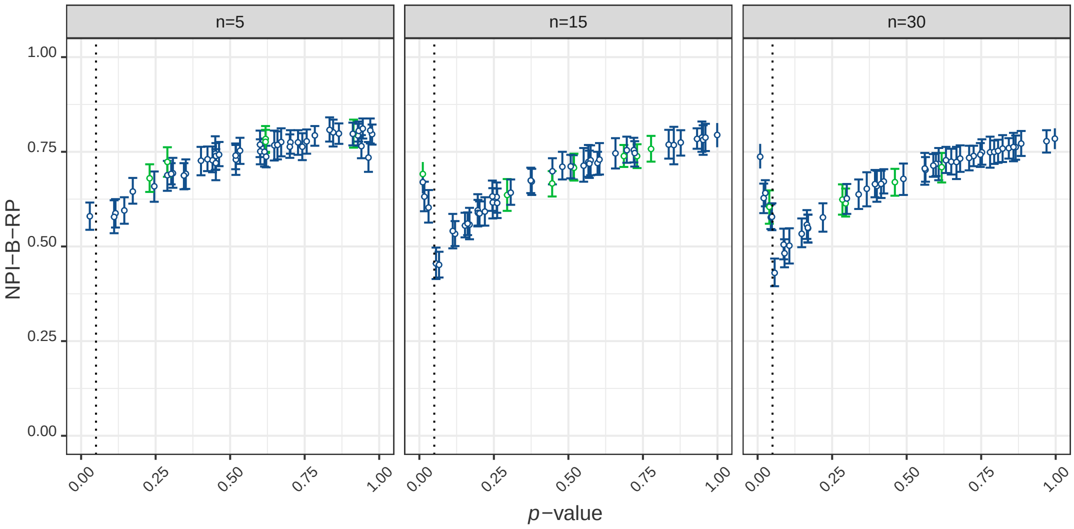


Figure 6.19: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0, 1)$, $M = 3$.

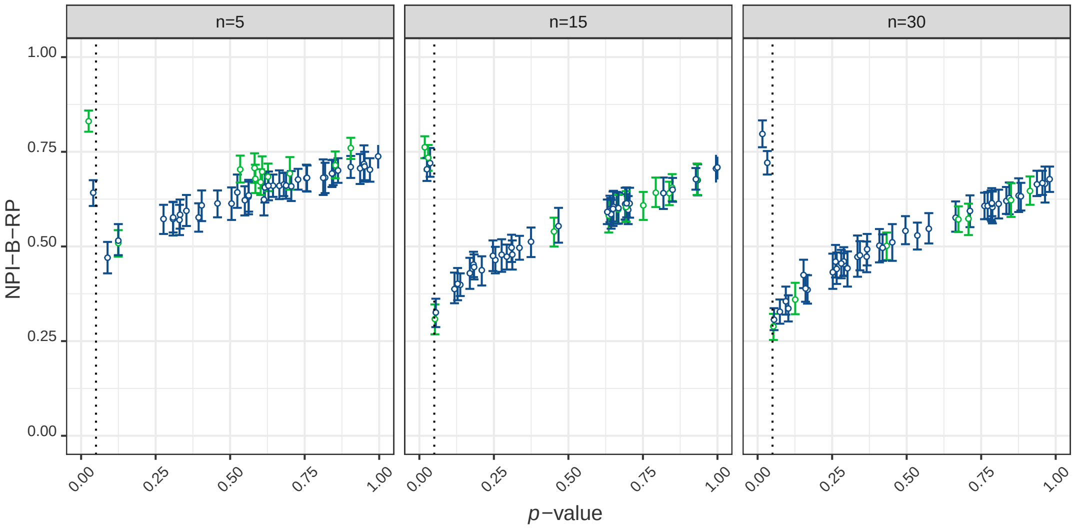


Figure 6.20: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0, 1)$, $M = 5$.

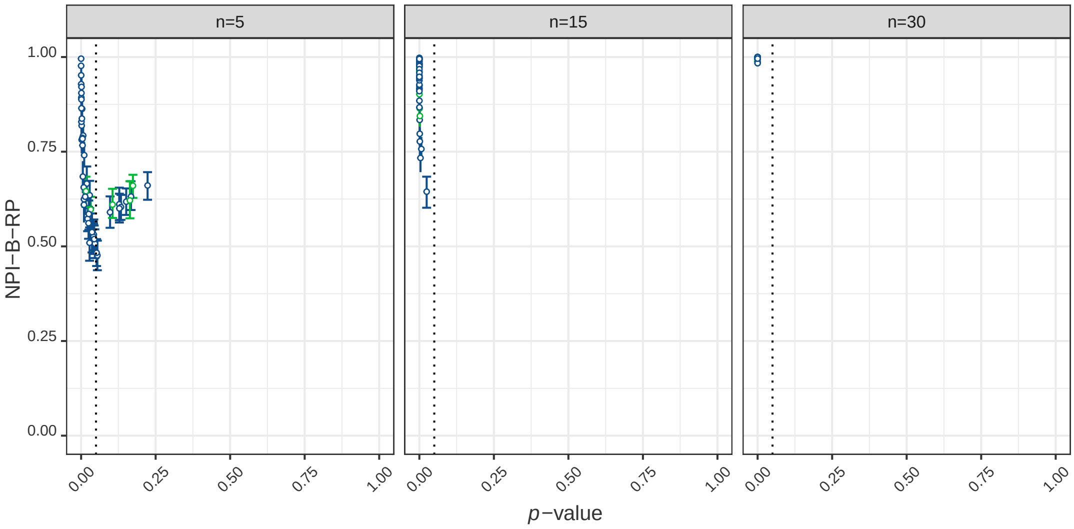


Figure 6.21: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

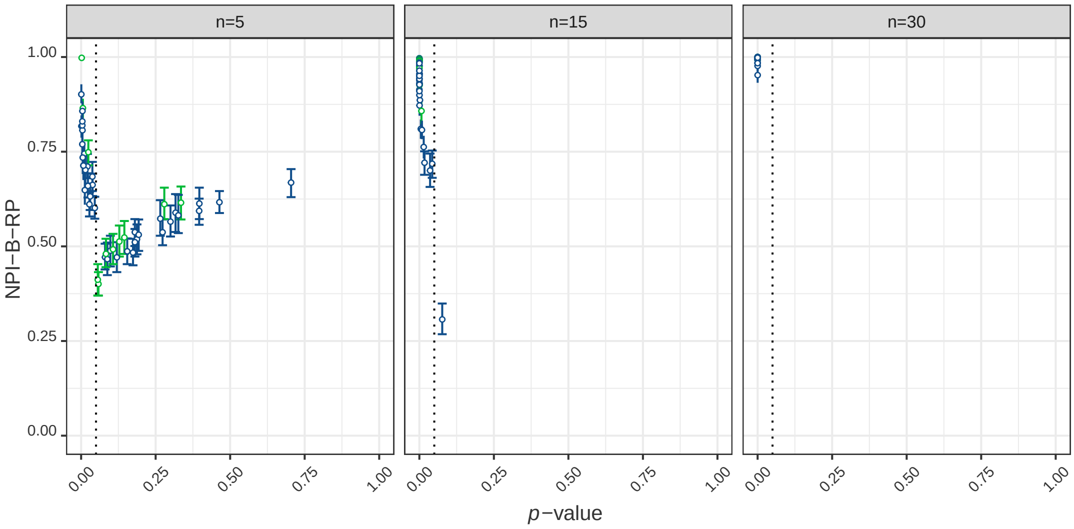


Figure 6.22: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(1,1)$, $N(0,1)$, $N(2,1)$,$N(2,2)$, and $N(0,2)$, $M = 5$.

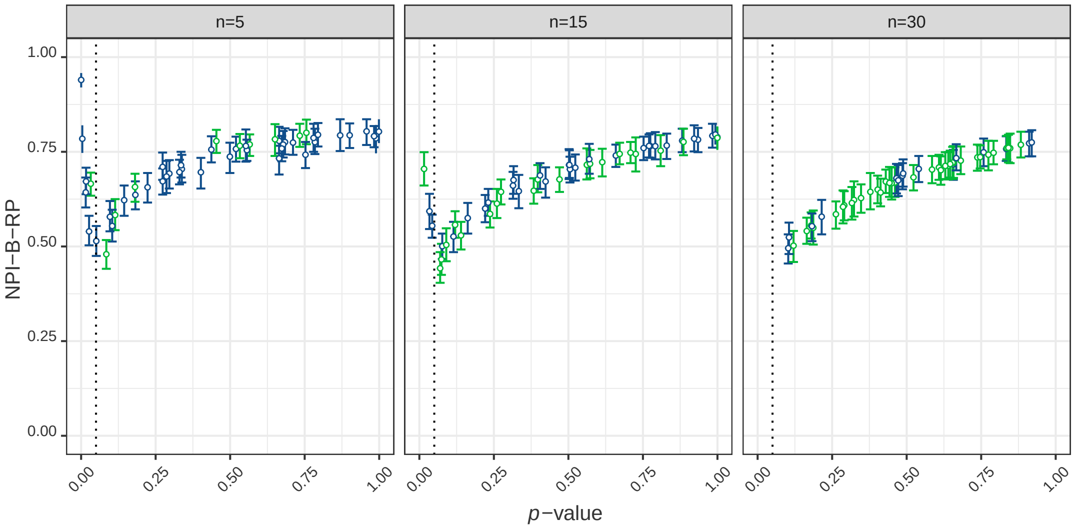


Figure 6.23: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $t(4)$, $M = 3$.

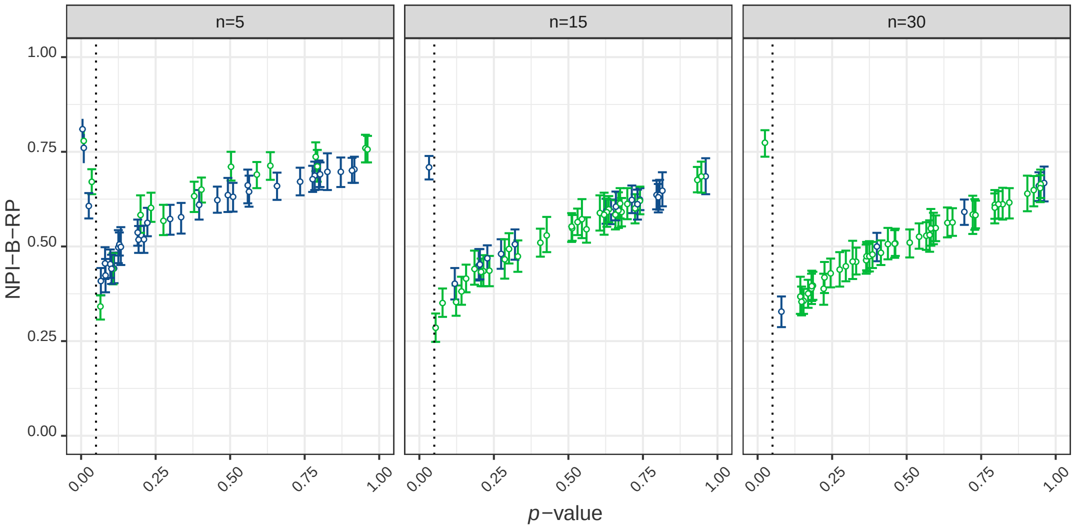


Figure 6.24: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $t(4)$, $M = 5$.

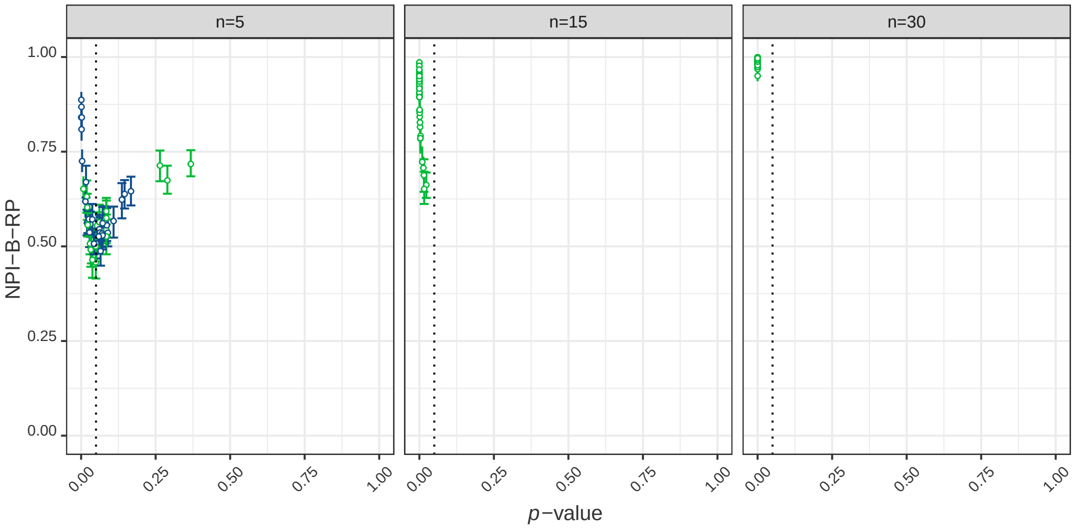


Figure 6.25: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0, 1)$, $N(1, 1)$, and $LN(1, 2)$, $M = 3$.

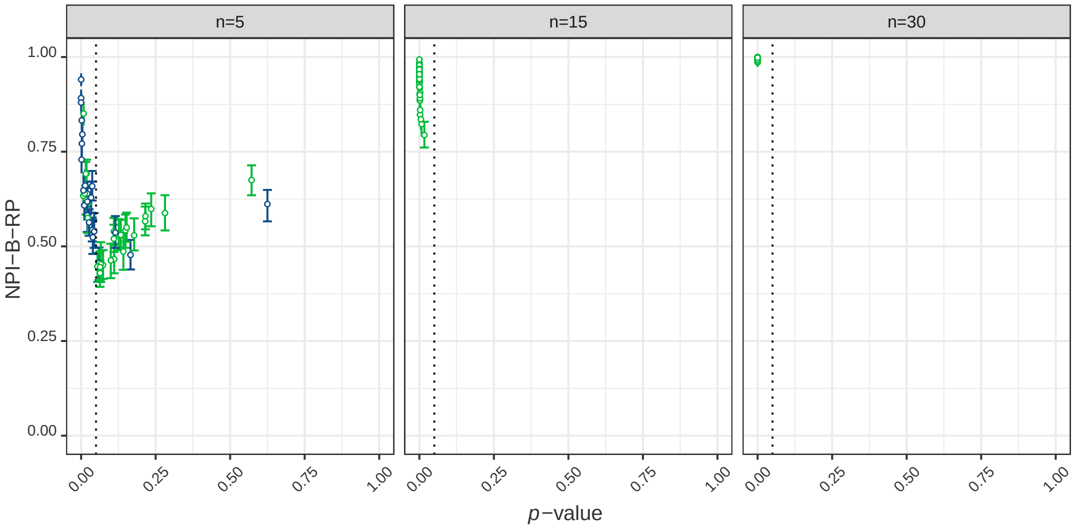


Figure 6.26: RP values for then two-stage procedure *Case C* against p-values for location test. ANOVA test (blue) and Kruskal-Wallis test (green). Samples are drawn from $N(0, 1)$, $N(1, 1)$, $LN(1, 2)$, $Ca(1, 1)$, and $t(4)$, $M = 5$.

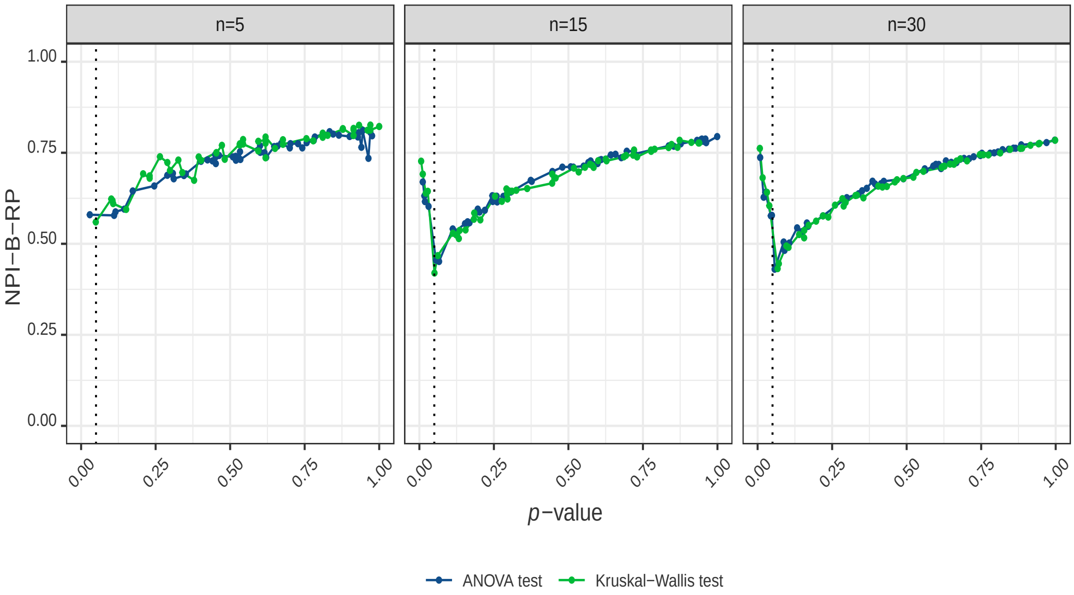


Figure 6.27: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $N(0, 1)$, $M = 3$.

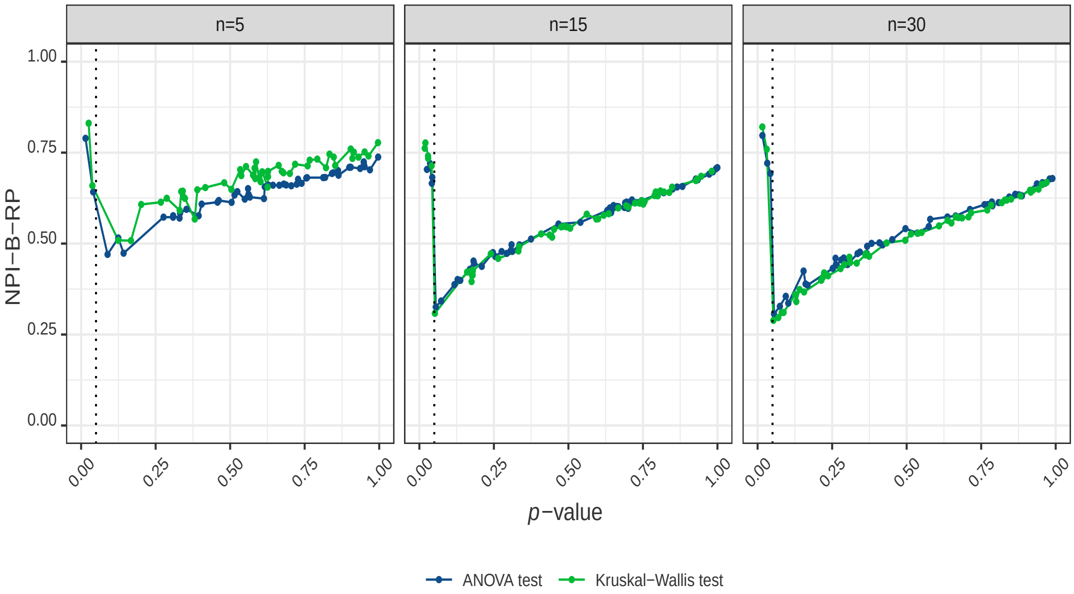


Figure 6.28: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $N(0, 1)$, $M = 5$.

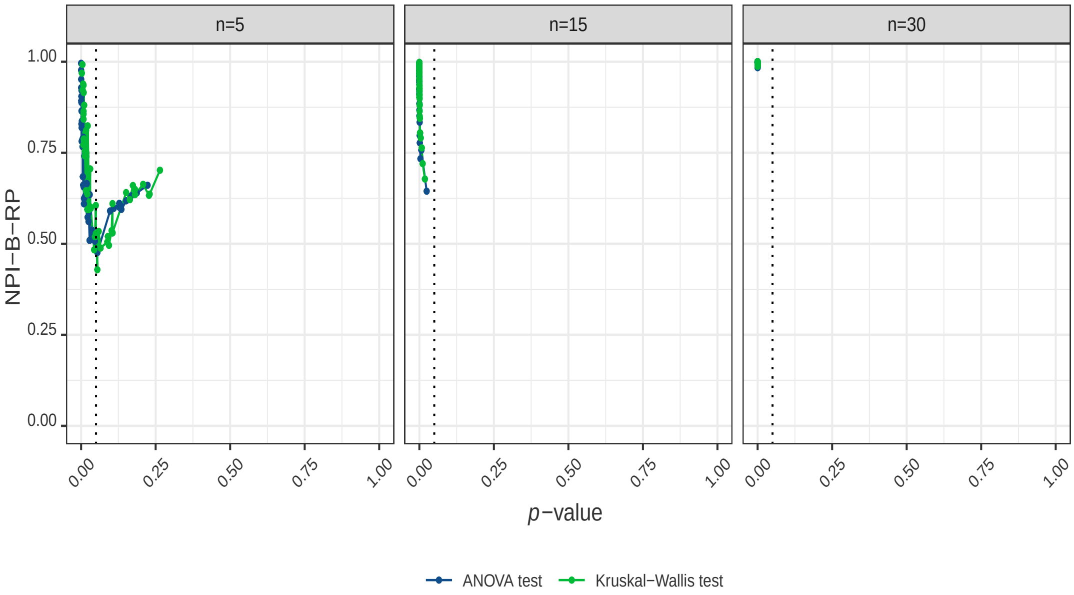


Figure 6.29: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

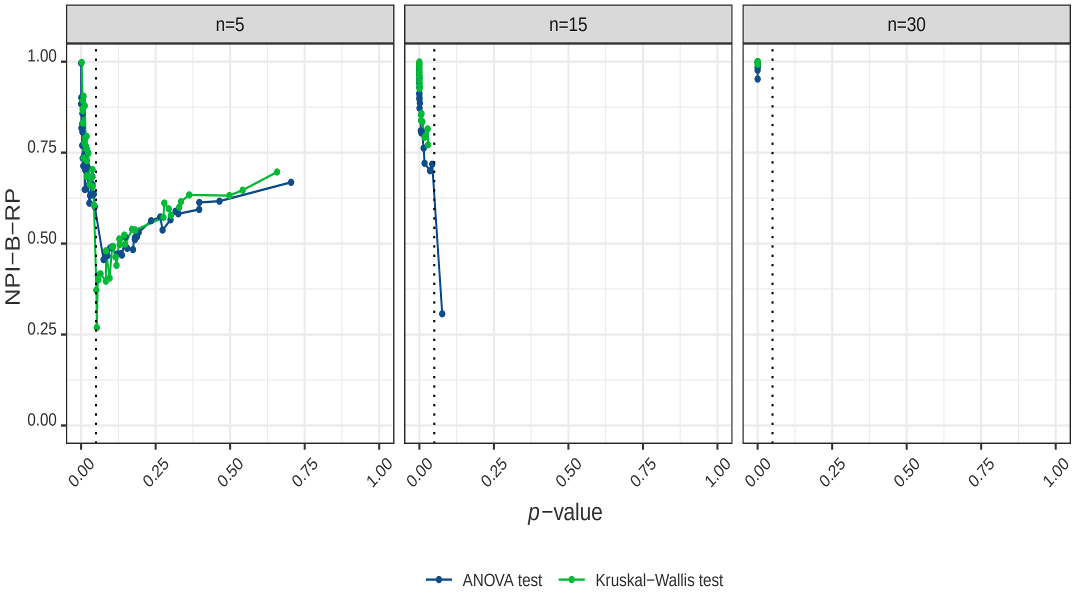


Figure 6.30: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $N(1,1)$, $N(0,1)$, $N(2,1)$, $N(2,2)$, and $N(0,2)$, $M = 5$.

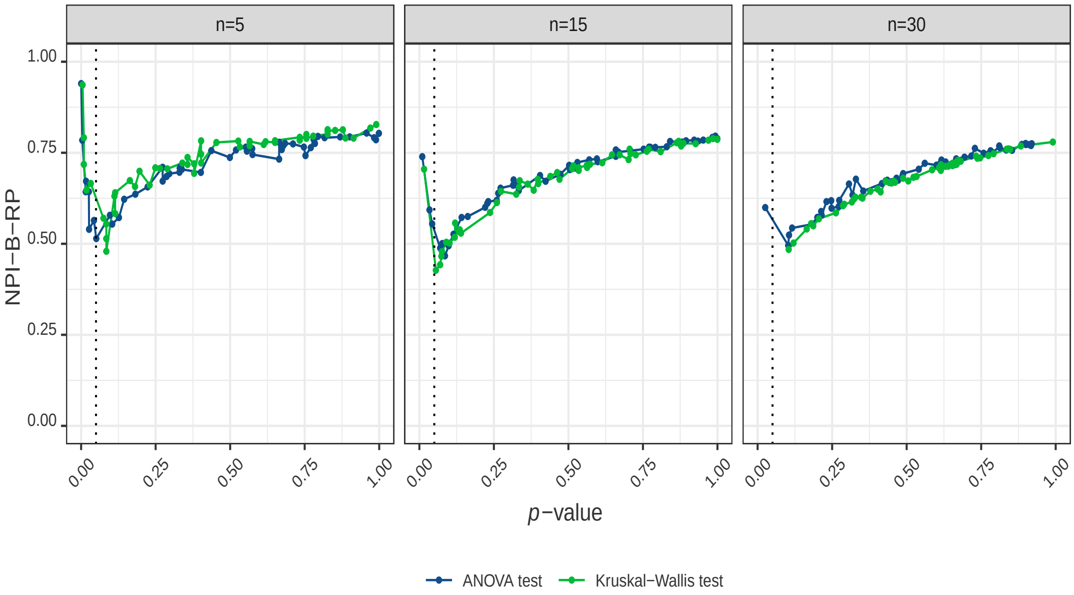


Figure 6.31: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $t(4)$, $M = 3$.

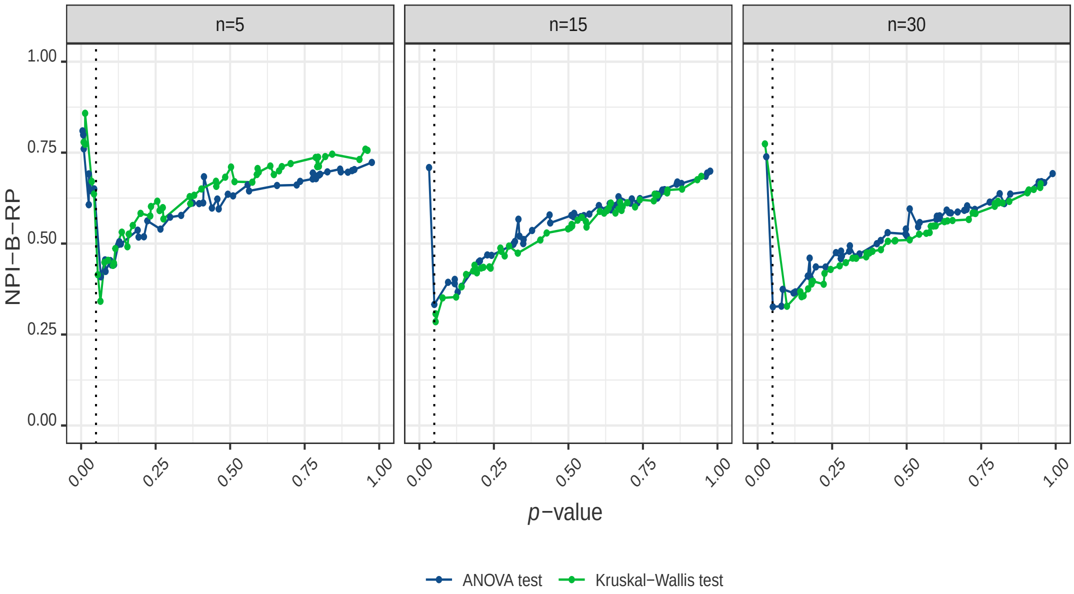


Figure 6.32: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $t(4)$, $M = 5$.

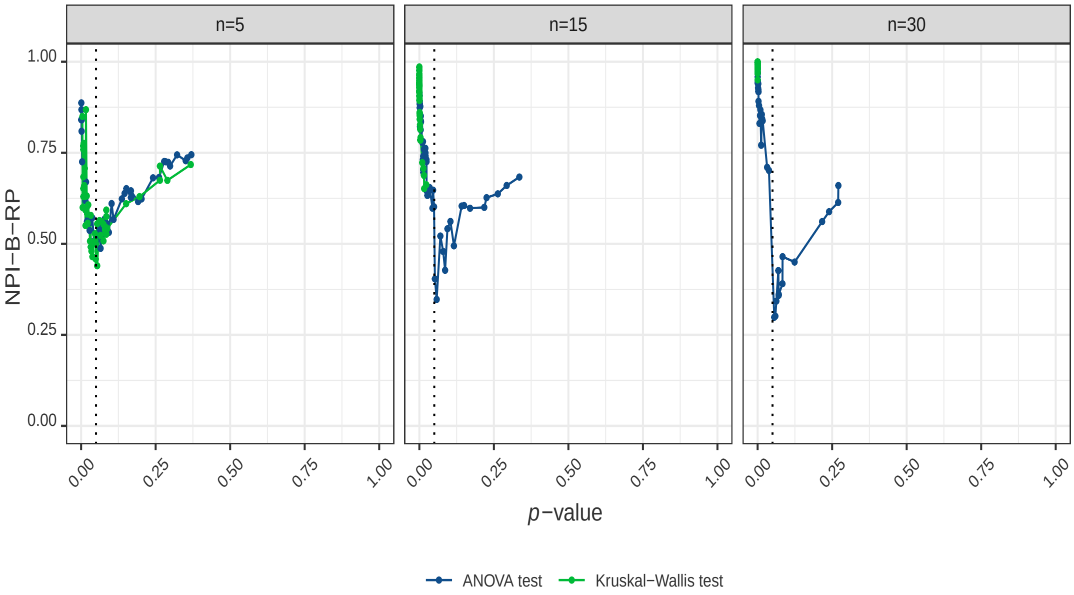


Figure 6.33: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $N(0, 1)$, $N(1, 1)$, and $LN(1, 2)$, $M = 3$.

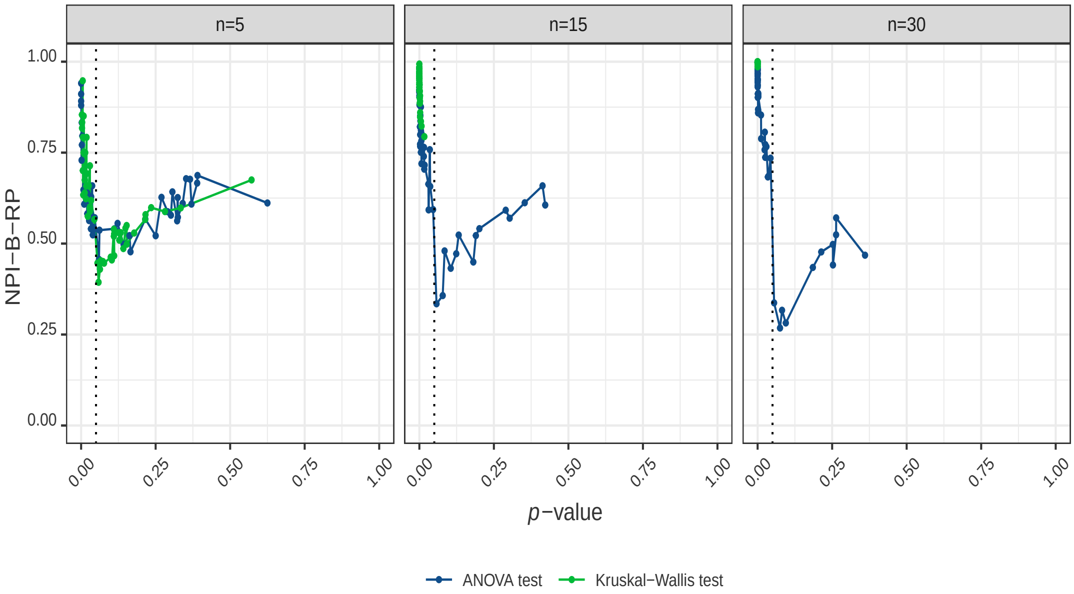


Figure 6.34: The means of RP values for the location tests without preliminary test against their p-values, samples are sampled from $N(0, 1)$, $N(1, 1)$, and $LN(1, 2)$,$Ca(1, 1)$, and $t(4)$, $M = 5$.

6.4 The impact of the preliminary test on the reproducibility of location tests

This section evaluates the effect of the preliminary test for Normality on the reproducibility of multiple-group location tests. This evaluation involves comparing the RP for the location test with the Normality test to the RP for the location test without the Normality test, as discussed in Section 4.4.

The Figures in this section illustrate the RP of location tests, with circles representing the ANOVA test and squares representing the Kruskal-Wallis test. The figures show two conditions: one with the preliminary test of Normality (represented by unfilled symbols) and the other without the preliminary test (filled symbols). The x-axis represents the mean of the p-values for location tests, while the y-axis represents the overall mean of RP values.

The results of the comparison for 5 groups are presented in Appendix E.1 because they yield the same conclusion as those for 3 groups.

6.4.1 The impact of the preliminary test on RP for *Case A*

The results of comparing full RP for the two stages with the product of the individual RPs for location tests and the preliminary test for 3 groups are presented, illustrating the impact of applying the preliminary Normality test on the RP values for the location tests. We compare the overall mean of RP values for location tests with the preliminary test (*Case A*) to the product of the overall mean of individual RP values for location and preliminary tests. Figures 6.35 - 6.38 show this comparison under different investigated distributions. There is no substantial difference between RP for the two-stage procedure for *Case A* and the product of the individual RP for location tests and the Normality test. However, the RP value for the Kruskal-Wallis test with the Normality test is slightly smaller than the product of individual RPs of Kruskal-Wallis and the Normality test. This difference decreases with increasing sample sizes. This is because the two-stage procedure for the three-sample results in increased error rates from both the Normality tests and the Kruskal-Walllis test. If any of the NPI-B samples fail incorrectly to pass the Normality test

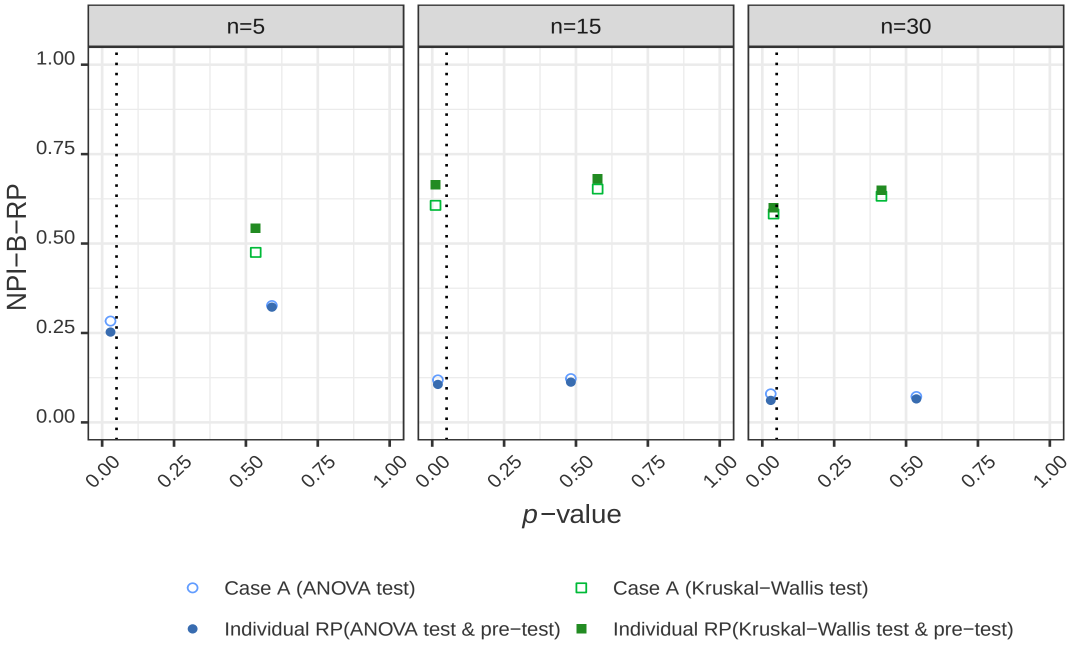


Figure 6.35: Comparison of the mean of RP for location tests with and without the preliminary test for *Case A* against the mean of their p-values, samples from $N(0, 1)$, $M = 3$.

this may lead to the unnecessary application of the Kruskal-Wallis test. Error aggravation can result in a lower overall RP compared to the product of individual RPs, as each test is considered independently. Thus, the reproducibility of the location tests, conditional on the outcomes of the Normality test is not substantially better than that of the location test alone.

6.4.2 The impact of the preliminary test on RP of location tests for *Case B*

The results of comparing the reproducibility of the outcome of the location tests (*Case B*) with the reproducibility of location tests without the preliminary test of Normality are presented.

Figure 6.39 illustrates the comparison between RP for location tests with and without the Normality test under the null hypothesis for both stages (Normality and equality of means) for 3 groups. Generally, the influence of the Normality test on RP for location tests is negligible. When considering small sample sizes, the RP for Kruskal-Wallis with

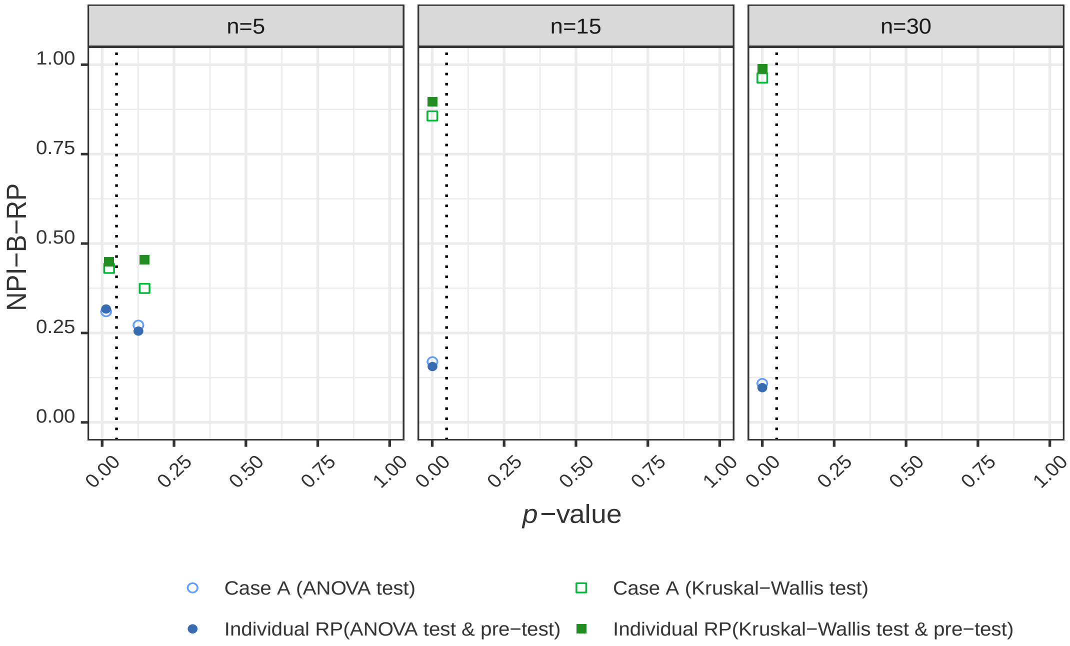


Figure 6.36: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

the preliminary test for Normality is lower than the RP for Kruskal-Wallis without the preliminary test in both areas. Conversely, the RP for the ANOVA test with the preliminary test is slightly higher than the RP for ANOVA without the preliminary test in the non-rejection area and lower in the rejection area. However, with larger sample sizes, the RP for Kruskal-Wallis with and without the preliminary test for Normality tends to converge to approximately the same value in both areas. On the other hand, the RP for the ANOVA test with the preliminary test becomes slightly smaller than the RP for ANOVA without the preliminary test in the non-rejection area and higher in the rejection area.

Figure 6.40 shows the comparison between RP of location tests with and without the preliminary test for Normality, under H_0^N and H_1^2, for 3 groups. There is a small difference between RP for location tests with and without the preliminary test. For the sample size of 5, the RP of the ANOVA test with the preliminary test is smaller than that without the preliminary test in the rejection area, and the RP of Kruskal-Wallis with the preliminary test is higher than that without the preliminary test, the opposite happens

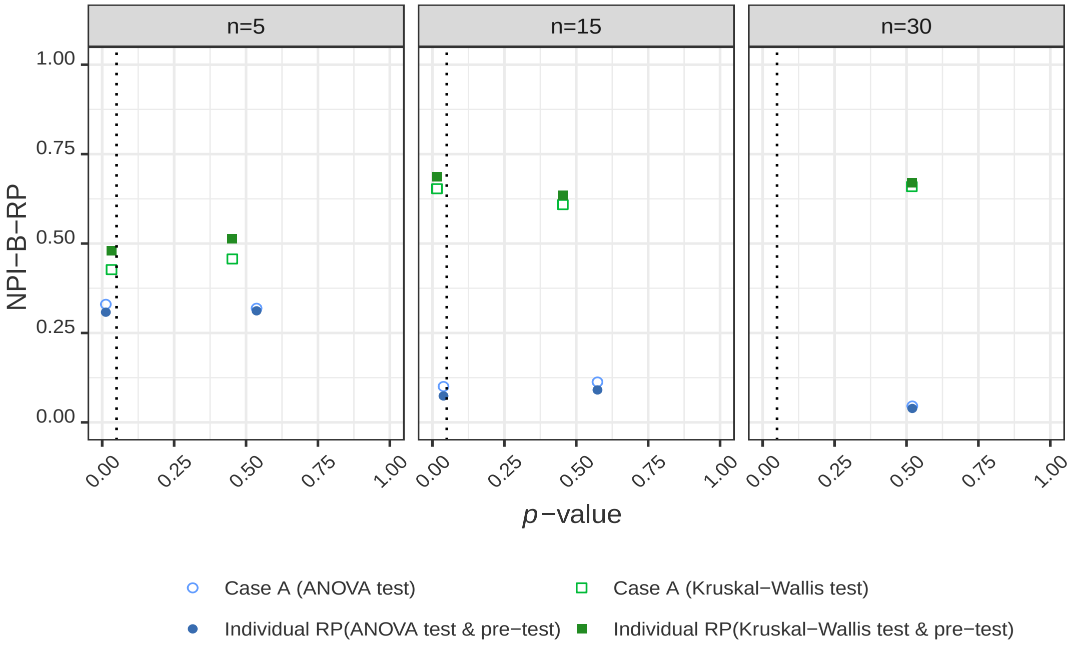


Figure 6.37: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $t(4)$, $M = 3$.

in the non-rejection area. However, this difference disappears as the sample sizes increase when applying the Kruskal-Wallis test.

Figure 6.41 shows the comparison between RP of location tests with and without the preliminary test for Normality, under H_1^N and H_0^2, for 3 groups. The impact of the Normality test on RP for location tests is negligible. For large sample sizes, there is no difference between the RP of Kruskal-Wallis with and without the preliminary test. However, for small sample sizes, it appears that the RP of Kruskal-Wallis with the preliminary test is slightly lower than the RP without the preliminary test in the non-rejection area. In the case of the ANOVA test, for the sample size of 15, RP with the preliminary test is slightly smaller than RP without the preliminary test in both areas.

Figure 6.42 shows the comparison between RP of location tests with and without the preliminary test for Normality, under H_1^N and H_1^2, for 3 groups. There is a small difference between RP for location tests with and without the preliminary test in the sample size of 5, generally, RP of both tests with the preliminary test tend to be smaller than that without the preliminary test. However, this difference disappears as the sample sizes

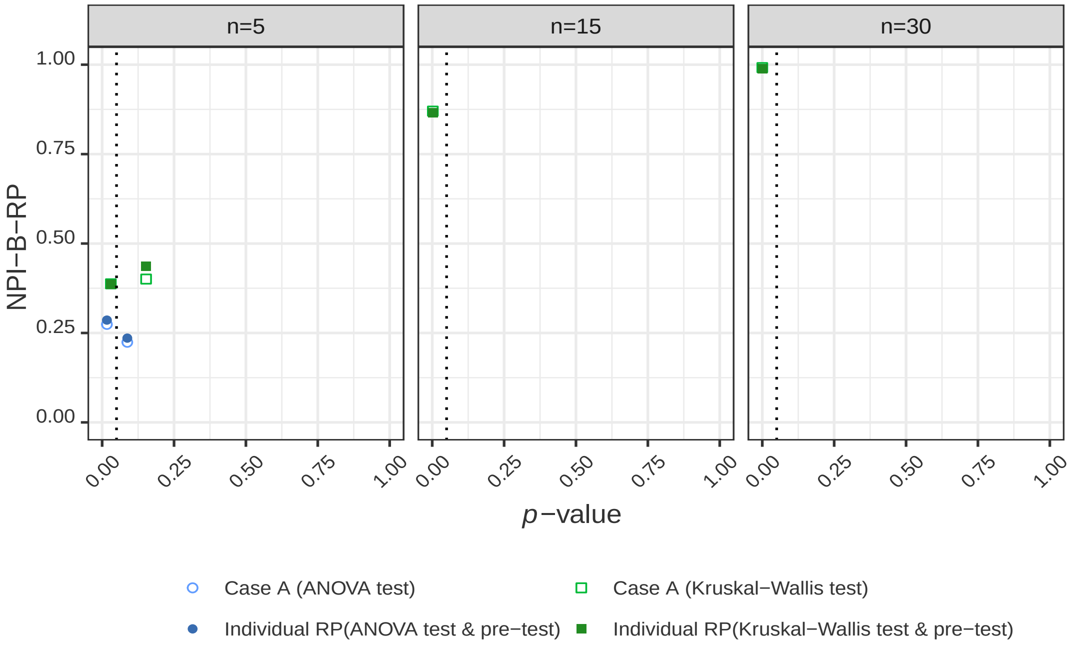


Figure 6.38: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $N(0,1)$, $N(1,1)$, and $LN(1,2)$, $M = 3$.

increase when applying the Kruskal-Wallis test.

Generally, the impact of the Normality test on the reproducibility of the outcome of location tests is minimal. RP values of the ANOVA test are more affected by the preliminary test of Normality than the Kruskal-Wallis test across most scenarios. With larger sample sizes, the influence of the preliminary test tends to diminish, and in some cases, it becomes negligible. Thus, the preliminary test for the Normality does not substantially enhance the reproducibility of the location test outcomes.

6.4.3 The impact of the preliminary test on RP for location tests for *Case C*

The comparison results for the reproducibility for *Case C* and reproducibility for location tests without preliminary tests are shown. By comparing the overall mean of RP values for the ANOVA test after filtering the original samples based on passing the Normality test with the overall mean of RP values for the ANOVA test without the Normality test for

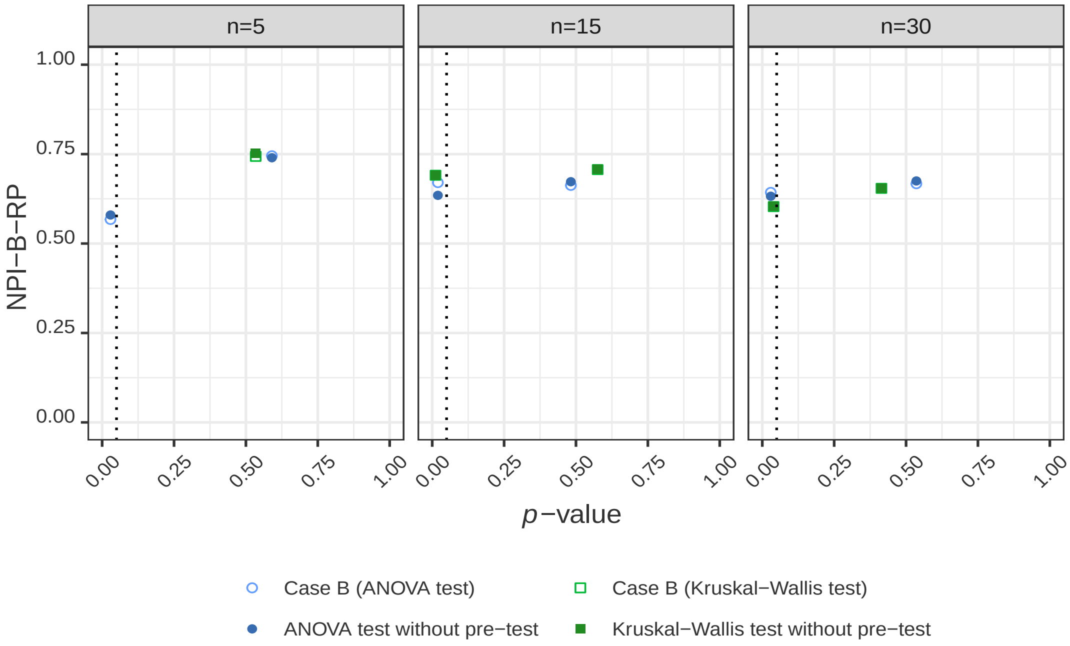


Figure 6.39: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $N(0, 1)$, $M = 3$.

all original samples. Likewise, comparing the overall mean of RP values for the Kruskal-Wallis test after filtering the original samples based on failure to pass the Normality test with the overall mean of RP values for the Kruskal-Wallis test without the Normality test for all original samples.

Figure 6.43 shows the comparison under H_0^N and H_0^2, it is evident that the effect of the preliminary test of Normality on RP of location tests is small. It seems that the RP of the ANOVA test has the least impact from applying the Normality test compared to RP for the Kruskal-Wallis test.

Figure 6.44 presents the comparison results under H_0^N and H_1^2. The ANOVA test appears to be minimally affected by the preliminary Normality test. However, for the Kruskal-Wallis test, RP with the preliminary test is better than RP without the preliminary test in the sample size of $n = 5$ in both areas. In contrast, for large sample sizes, there is little difference between the RP values for the Kruskal-Wallis test with and without the preliminary test.

Figure 6.45 illustrates the result of comparison when all original samples are not

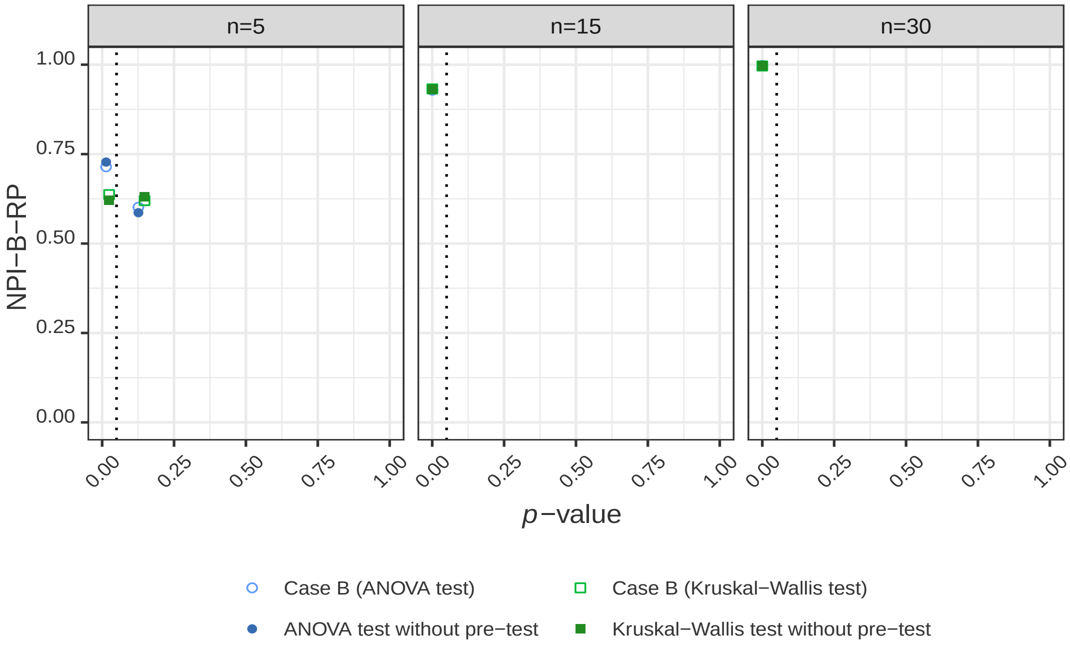


Figure 6.40: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $N(1, 1)$, $N(0, 1)$, and $N(2, 1)$, $M = 3$.

Normal and have the same mean for 3 groups. The effect of the preliminary test for Normality on RP of location tests is small. For sample sizes of 5 and 15, there are slight differences between RP of location tests with and without the Normality test. For the sample size of 30, there is no difference between RP of location tests with and without the Normality test.

Figure 6.46 displays the comparison results under H_1^N and H_1^2. We observe that for the Kruskal-Wallis test, the preliminary Normality test does not substantially affect RP values in sample sizes of 15 and 30. However, for the sample size of 5, the RP for the Kruskal-Wallis test without the preliminary test is higher than that with the preliminary test in the rejection area. In the case of ANOVA, there is no substantial difference in RP values in the rejection area between tests with and without the preliminary test in the sample size of 5. However, in the non-rejection area, RP for ANOVA with the preliminary test is lower than that without the preliminary test.

Generally, the impact of filtering the original samples based on the Normality test

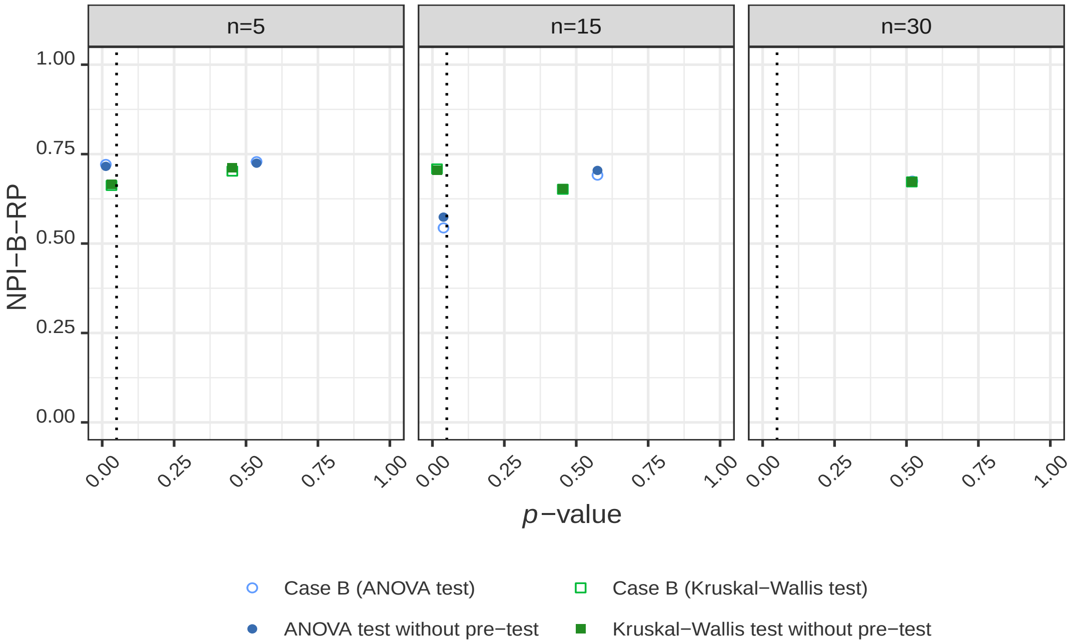


Figure 6.41: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $t(4)$, $M = 3$.

results on RP of location tests is small. As sample sizes increase, this impact tends to diminish and, in some instances, becomes negligible. Thus, filtering the original samples based on the Normality test result does not substantially improve RP for location tests without preliminary tests.

Similar results were also achieved when the simulation was conducted for bimodal distribution from a mixture of Normal distributions and the results are listed in Appendix E.2.

A brief examination of RP for the three-stage procedure was conducted, which involved the preliminary Normality test (Shapiro-Wilk) and the preliminary test of equality of

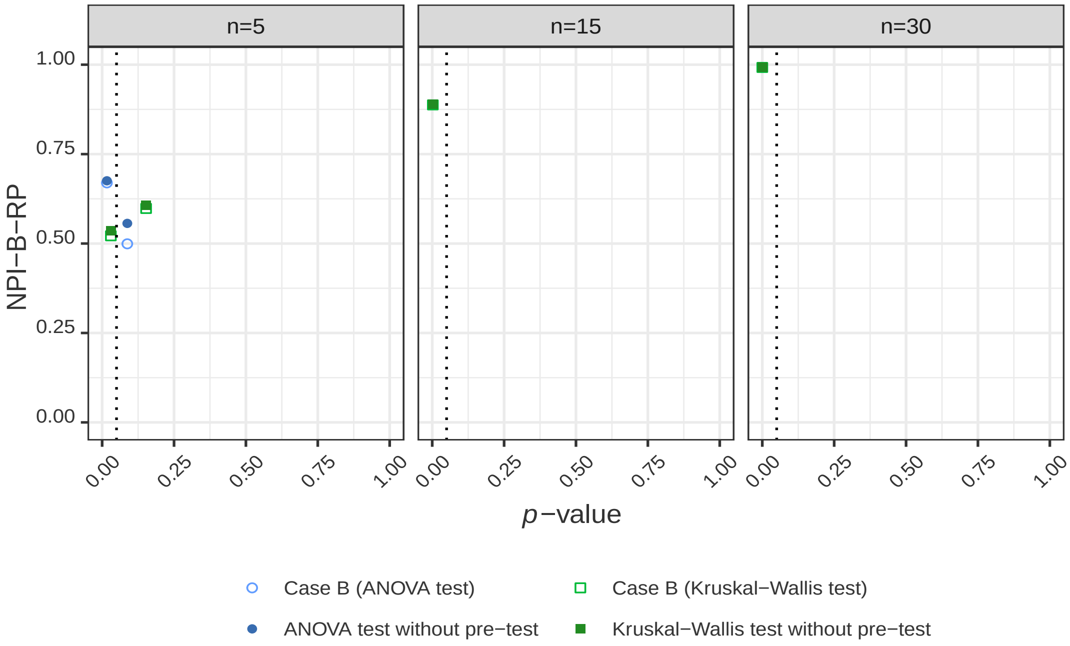


Figure 6.42: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $N(0,1)$, $N(1,1)$, and $LN(1,2)$, $M = 3$.

variances (Levene's test), followed by the location test (one-way ANOVA, one-way Welch's ANOVA, or Kruskal-Wallis) based on the outcomes of the preliminary tests. The same results as observed in the two-stage procedure are obtained, this can be found in Appendix E.3

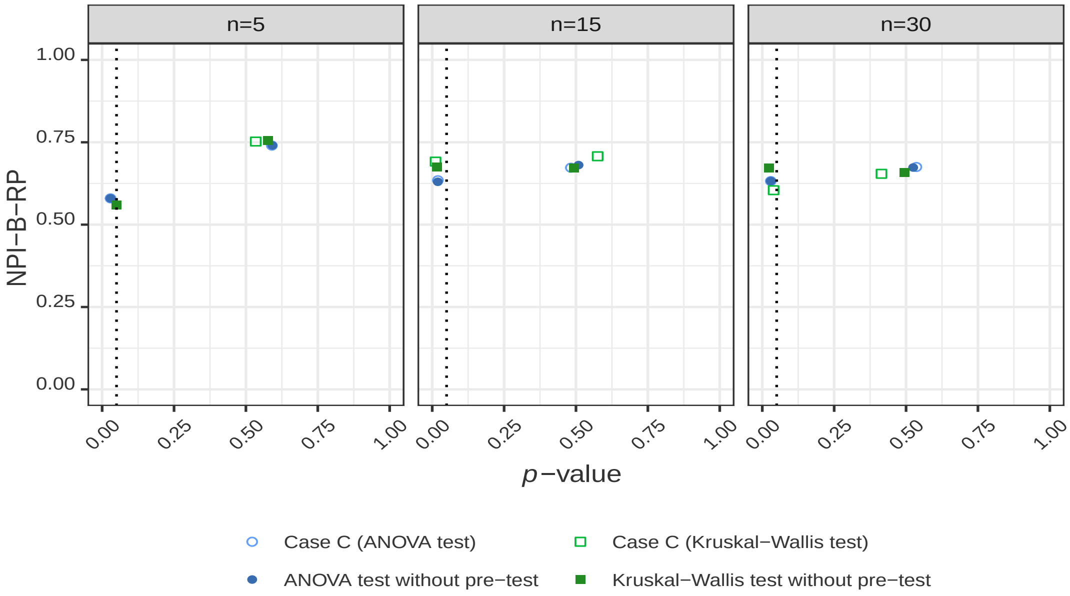


Figure 6.43: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $N(0,1)$, $M = 3$.

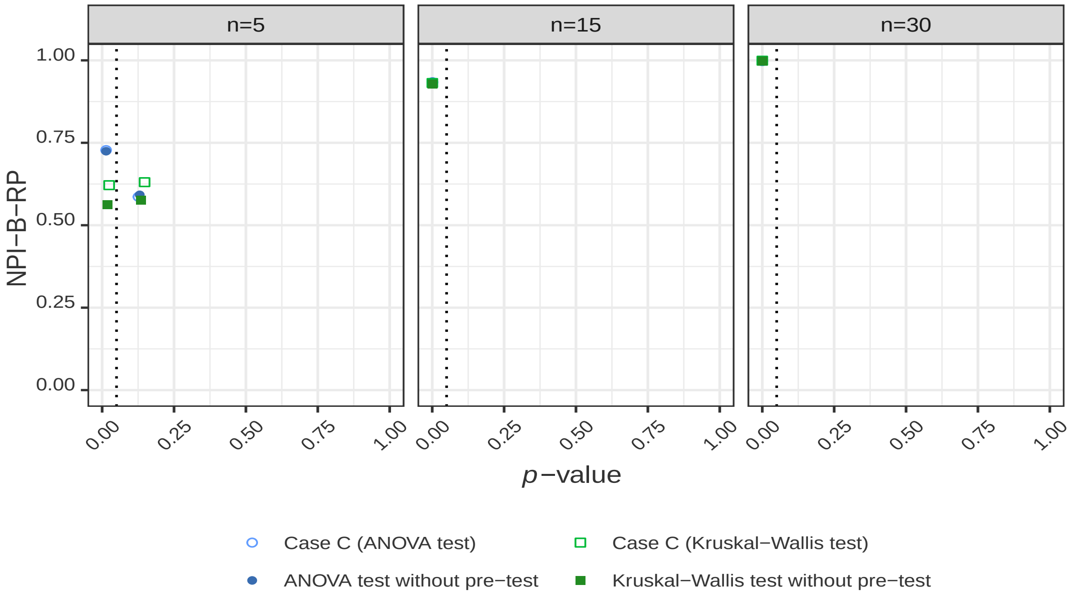


Figure 6.44: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $N(1,1)$, $N(0,1)$, and $N(2,1)$, $M = 3$.

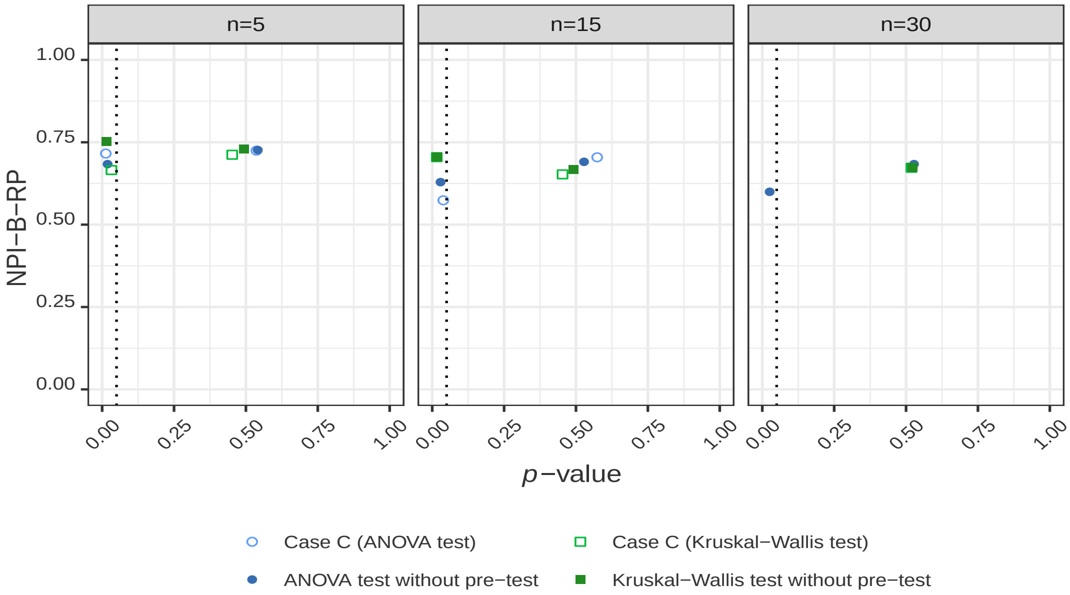


Figure 6.45: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $t(4)$, $M = 3$.

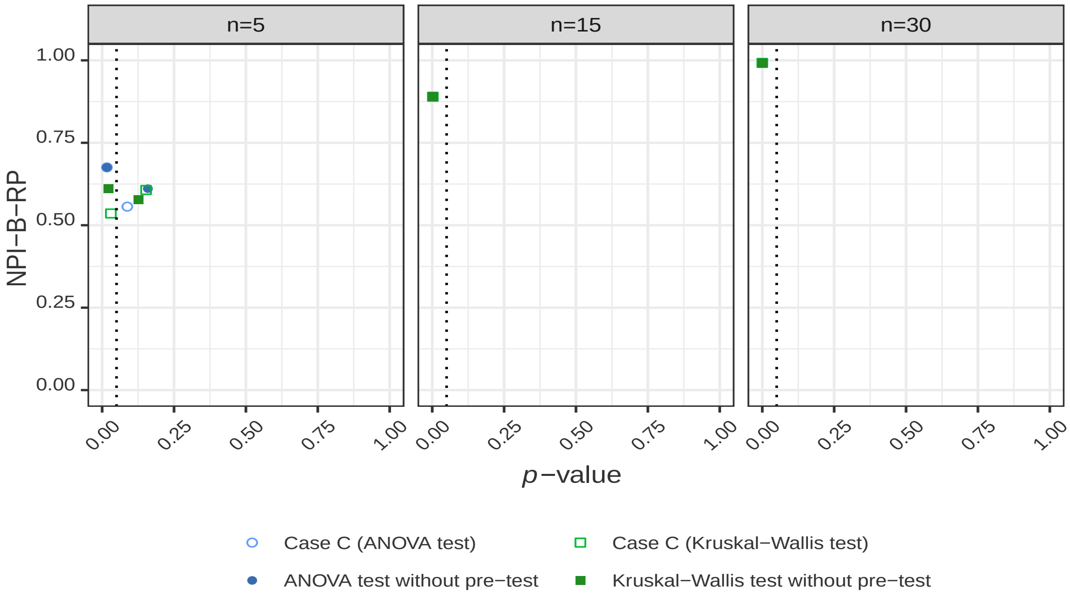


Figure 6.46: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $N(0,1)$, $N(1,1)$, and $LN(1,2)$, $M = 3$.

6.5 Conclusion

This chapter has studied the reproducibility probability (RP) for the two-stage procedure for multiple groups. The two-stage procedure involves the preliminary Normality test (Shapiro-Wilk), followed by the location test (one-way ANOVA or Kruskal-Wallis) based on the outcome of the Normality test. Additionally, this chapter explores the reproducibility of multiple-group location tests without the preliminary test. The primary focus was on investigating the impact of the Normality test on RP for multiple-group location tests.

The RP of this two-stage procedure was examined under various Cases. The first Case is *Case A* which studies full RP for two stages. Then, *Case B* focuses on the RP of the outcome of the location tests. Finally, *Case C* studies RP of the location test conclusion, where for the NPI-B samples the same location test is applied as for the original sample.

Simulation studies are conducted to investigate the aims of this chapter. The results of the simulation show that RP for multiple-group location tests with and without preliminary tests show the general pattern of RP: RP values are low when p-values for location tests are close to the threshold, and RP are high when p-value for location tests is far away from the threshold. RP values for parametric tests without preliminary tests tend to be higher than nonparametric tests in the non-rejection area when the sample size is large, and vice versa in the small sample sizes.

The chapter also examined the influence of sample size on the RP of location tests with and without the Normality test. It was observed that, with larger sample sizes, the RP of location tests tends to decrease in the non-rejection area. Conversely, in the rejection area, the RP tends to increase as the sample size increases. In *Case A*, this impact is true for the Kruskal-Wallis test, but for RP values of the ANOVA test are decreased in both areas.

Furthermore, this chapter demonstrated the impact of the number of groups on RP values. As the number of groups increases, RP values for location tests typically decrease in the non-rejection area while increasing in the rejection area. The only exception to this trend is observed in *Case A* of the RP for the two-stage procedure, where RP values for the ANOVA test decrease in both areas as the number of groups increases.

When comparing the RP for the location tests with and without the Normality test, there is no significant difference between RP for the multiple-group location tests with and without the Normality test, which means that the influence of the preliminary test on RP of location tests is small. Moreover, the findings revealed that with larger sample sizes, the influence of the preliminary tests on RP of the location tests tends to diminish, and in certain scenarios, it becomes negligible.

Additionally, this chapter examined the relationship between the overall mean of RP values in the rejection area and the estimated power for location tests with and without preliminary tests for 3 groups. This relationship is positive, the RP increases as the power increases. RP values and the estimated power for the Kruskal-Wallis test are better than ANOVA when dealing with non-Normal distributions and large sample sizes.

Chapter 7

Conclusions and Future Work

7.1 Conclusions

In conclusion, this thesis has investigated the impact of applying preliminary tests on the reproducibility probability (RP) of location tests. Through applying simulation studies, we have explored the reproducibility of location tests with and without preliminary tests and the extent of the impact of conducting preliminary tests on RP of location tests. Our investigation revealed that the effect of preliminary tests on the RP of location tests is small, that is applying preliminary tests does not lead to a substantial improvement or deterioration in the reproducibility of the location tests.

Chapter 6 primarily investigates the use of preliminary tests for assessing the assumptions of ANOVA. However, it is important to recognize that in practical applications, researchers often rely more heavily on residual diagnostics to evaluate model assumptions after fitting the model. If preliminary tests indicate Normality and homoscedasticity but residual diagnostics suggest otherwise, this discrepancy needs careful consideration. This discrepancy might highlight potential issues with the initial assumptions or indicate limitations in the sensitivity of preliminary tests, suggesting that preliminary tests might not always capture the complexity of the data or the nuances introduced by the model fitting process. Studying reproducibility taking such further model checking practices into account, beyond preliminary tests, is an important topic for future research.

The thesis began by studying the reproducibility of Normality tests in Chapter 2, which presented the reproducibility probability (RP) for three of the most famous Normality tests: the Shapiro-Wilk test, Anderson-Darling test, and Lilliefors test, and compared their performance in terms of RP. Moreover, their RP for different levels of significance was studied. The findings show that RP for Normality tests tends to be low when their p-value is close to the threshold, and RP increases gradually as the p-value moves away from the level of significance. RP values vary according to sample size: for small sample sizes, RP values in the non-rejection area tend to be very high, while RP in the rejection area is very small. As the sample size increases, RP in the non-rejection area decreases and RP in the rejection area increases. Overall, there is no substantial difference between the RP of these tests. However, RP values of the Anderson-Darling test have less variability than RP for the Shapiro-Wilk and Lilliefors tests. From the results, the Anderson-Darling test has the slightly highest RP in the rejection area, while the Shapiro-Wilk test has the highest RP in the non-rejection area. RP values for the Normality tests in the non-rejection area tend to be high when the significance level is low, and RP values in the rejection area tend to be low. As the significance level increases, RP in the non-rejection area decreases and RP in the rejection area increases.

Chapter 3 addressed the reproducibility of equality of variances tests. RP for the F-test and Levene's test were studied. The results show that both tests have high RP when their p-values are close to the level of significance, and as the p-value moves away from the level of significance, their RP increases. We found that RP for the two-sided F-test is not close to one if the p-value is close to one, while in the upper-sided F-test RP is very close to one as the p-value approaches one. Additionally, we observed that when dealing with non-Normal data, the RP values for Levene's test are better than RP for the F-test in terms of variability and RP value. The relationship between the overall mean of RP values in the rejection area and the estimated power of the equality of variances tests is positive; as the power of the test increases, the RP value increases. Similarly, for the relationship between the sample size and the power and RP, as the sample size increases, the power and RP increase.

Chapter 4 investigates the reproducibility of one-sample location tests, specifically focusing on the one-sample t-test and the one-sample Wilcoxon test. The examination encompasses scenarios with and without the inclusion of a preliminary test for Normality.

In the scenario of RP for location tests with a preliminary test (the two-stage procedure), the preliminary test for Normality is applied in the first stage. In the second stage, the t-test is performed if the null hypothesis for Normality is not rejected, otherwise, the Wilcoxon test is applied. Three cases of RP for the two-stage procedures were considered: *Case A*: full RP for all stages. *Case B*: RP for the outcome of the location test. *Case C* studies RP of the location test conclusion, where for the NPI-B samples the same location test is applied as for the original sample. RP results for location tests with and without the Normality test show the general pattern for RP: RP values are low when p-values of location tests are close to the threshold and higher RP values when the p-value is far away from the threshold. As the sample size increases, RP values in the non-rejection area tend to decrease while RP values in the rejection area increase. Full RP for all stages (*Case A*) shows that for a small sample size, RP for the t-test tends to be higher than RP for the Wilcoxon test; with increasing sample size, RP for the t-test decreases while RP for the Wilcoxon test increases. The impact of preliminary tests on RP of location tests demonstrated no substantial difference. The influence of preliminary tests on RP of location tests diminishes substantially for large sample sizes and non-Normally distributed data. Additionally, this chapter examined the relationship between the overall mean of RP values in the rejection area and the estimated power for location tests with and without the Normality test. This relationship shows that as the power of location tests increases the RP increases.

In Chapter 5, we investigated the RP of the two-sample location tests with and without preliminary tests. In the scenario of RP for location tests with preliminary tests (the three-stage procedure), location tests (t-test, Welch's t-test, and Wilcoxon-Mann-Whitney) are chosen according to the preliminary tests for Normality and equality of variances results. Findings from simulation studies revealed that RP values show the same results observed in Chapter 4. The impact of the preliminary tests of Normality and equality of variances on RP of two-sample location tests is small.

Chapter 6 investigates the RP for the two-stage and three-stage procedures for comparing means in multiple groups. The two-stage procedure involves a preliminary Normality test followed by the multiple-sample location tests one-way ANOVA test or Kruskal-Wallis test, while the three-stage procedure adds a preliminary test of equality of variances. The study was applied to three groups and five groups. *Cases A, B,* and *C* are examined, revealing that the impact of preliminary tests on RP for location tests is generally limited. The chapter also explores the influence of sample size and the number of groups on RP of location tests with and without preliminary tests, indicating that, with larger sample sizes, the impact of the preliminary test tends to diminish, and as the number of groups increases leads to decreased RP of location tests in the non-rejection area and increase RP in the rejection area. Moreover, this chapter examined the relationship between the overall mean of RP values in the rejection area and the estimated power for location tests with and without preliminary tests. this relationship shows that as the power of location tests increases the RP increases. RP and power for the Kruskal-Wallis test are better than ANOVA when dealing with non-Normal distributions and large sample sizes.

7.2 Future Work

The recommendations for the future work of this thesis are listed as follows:

1. This thesis investigated reproducibility for location tests and the preliminary tests under equal sample sizes. Exploring RP for location tests alongside preliminary tests for samples with varying sizes adds valuable dimensions to the research.

2. The research can be extended to multivariate tests, for example, multivariate t-tests and multivariate analysis of variance (MANOVA). Examine how the preliminary test of Normality impacts the RP of multivariate tests and compare the findings with those from univariate tests.

3. Using the Anderson-Darling (AD) test to assess the Normality assumption may lead to less variability in the results for the reproducibility probability for two and three-stage procedures.

4. Studying the impact of applying the preliminary test for symmetry on RP for the one-sample Wilcoxon signed-rank test, where the Wilcoxon test assumes the distribution of the paired differences is symmetric around the median. A preliminary test for symmetry such as the triples test proposed by Randles et al [76] can be

suggested to use before performing the Wilcoxon test. If the assumption is violated, then the sign test can be used.

5. While Chapter 6 primarily examines the impact of preliminary tests on ANOVA reproducibility, it is crucial to consider the influence of post-hoc tests on reproducibility as well. Commonly employed tests such as Tukey's Honestly Significant Difference (HSD) [1] and Bonferroni correction [15] are frequently used for pairwise comparisons following significant ANOVA results. These post-hoc tests are designed to control for Type I errors when making multiple comparisons. However, their conservative nature can potentially increase the risk of Type II errors, thereby affecting the reproducibility of research findings. Future research should study deeper into quantifying the impact of various post-hoc tests on reproducibility, and investigating how these factors influence the replicability of research outcomes across different experimental contexts and sample sizes.

Appendix A

Reproducibility for Normality Tests

A.1 Extended simulation examples

This section contains detailed results from the additional simulations involving data generated from a mixture of Normal distributions $X \sim 0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$. The results are presented in Figures A.1- A.4. The RP values for these cases are approximately similar to those from the main examples discussed in Chapter 2.

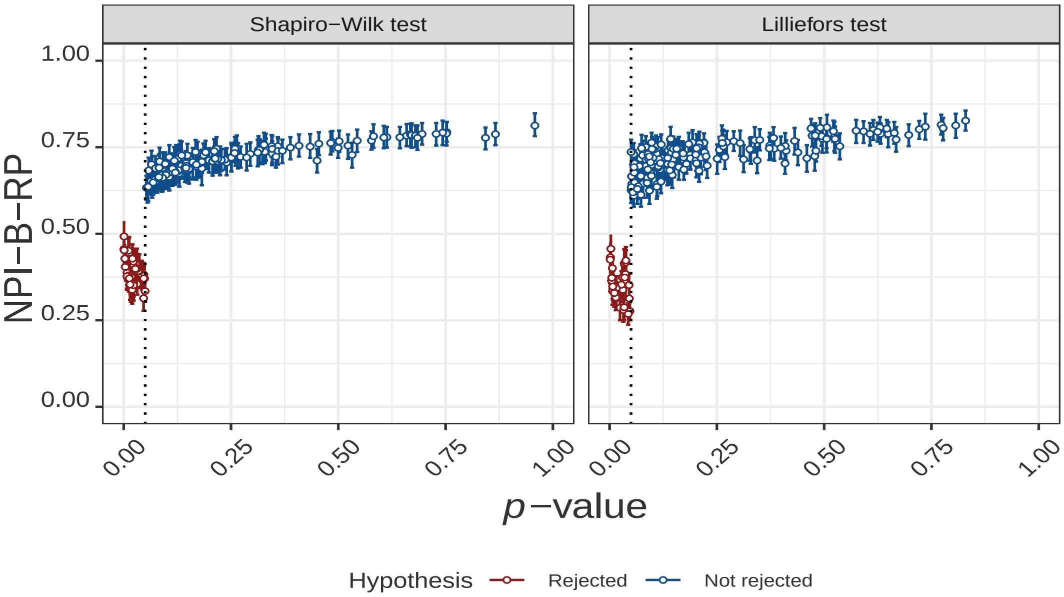


Figure A.1: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from a mixture of normal distributions $0.4 \cdot N(5,1^2)+0.6 \cdot N(15,2^2)$, with $n=5$ and $\alpha=0.05$

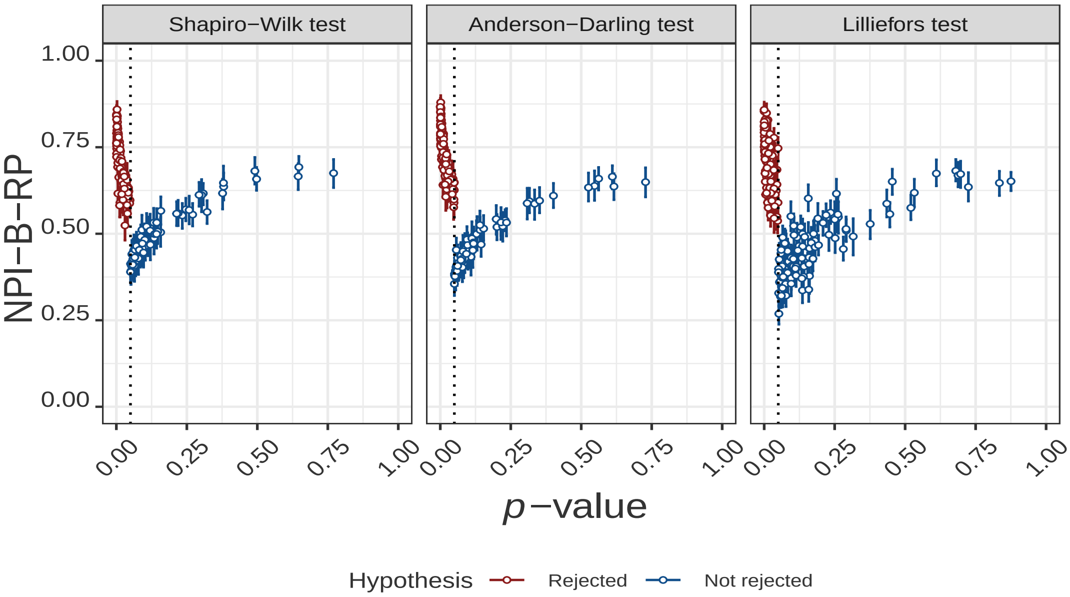


Figure A.2: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from a mixture of normal distributions $0.4 \cdot N(5,1^2)+0.6 \cdot N(15,2^2)$, with $n=10$ and $\alpha=0.05$

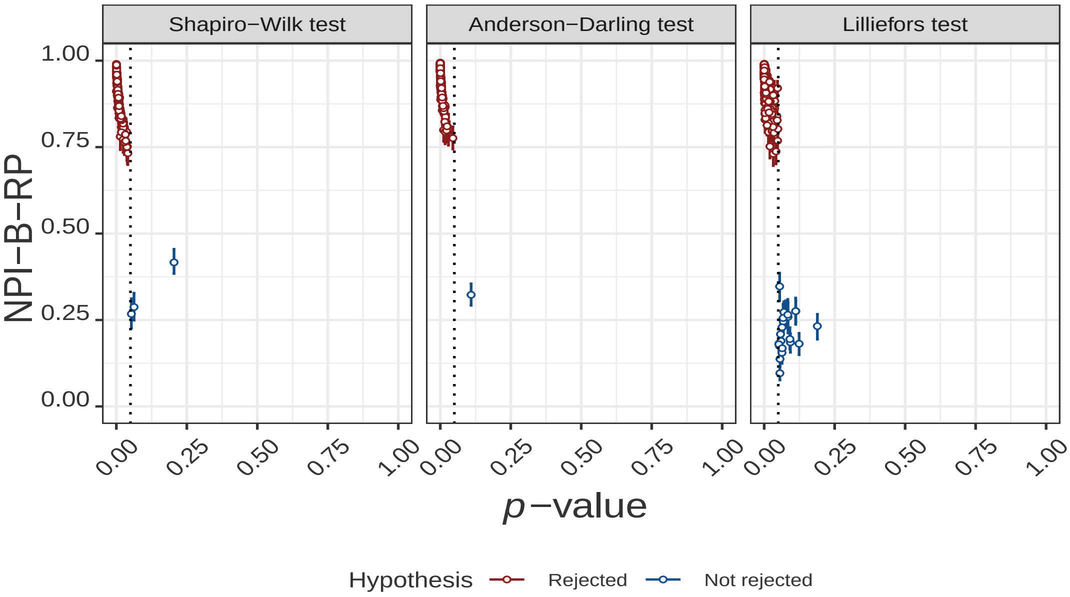


Figure A.3: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from a mixture of normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, with $n = 20$ and $\alpha = 0.05$

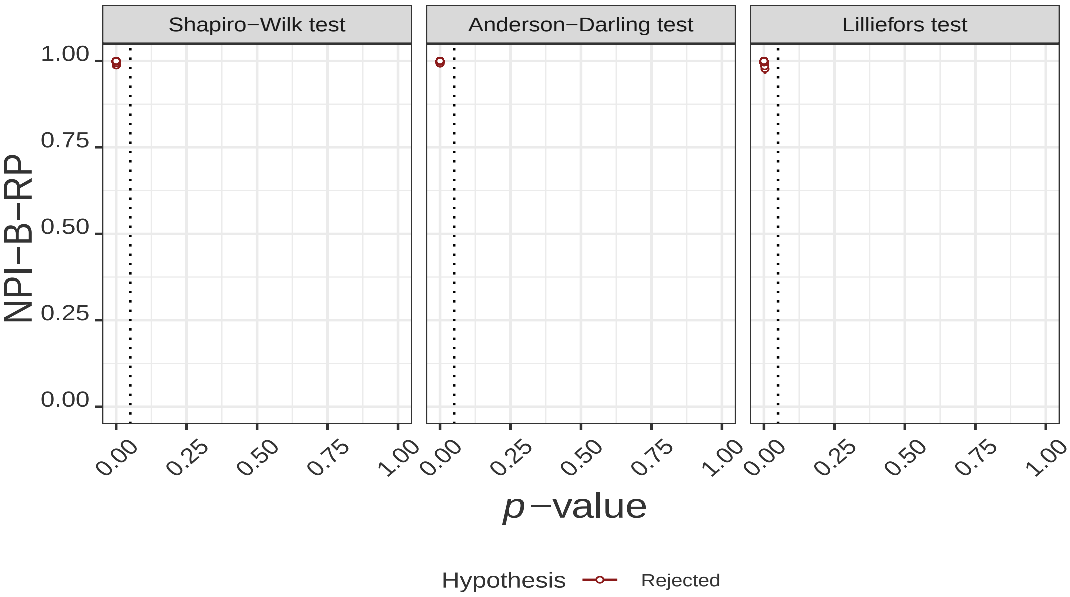
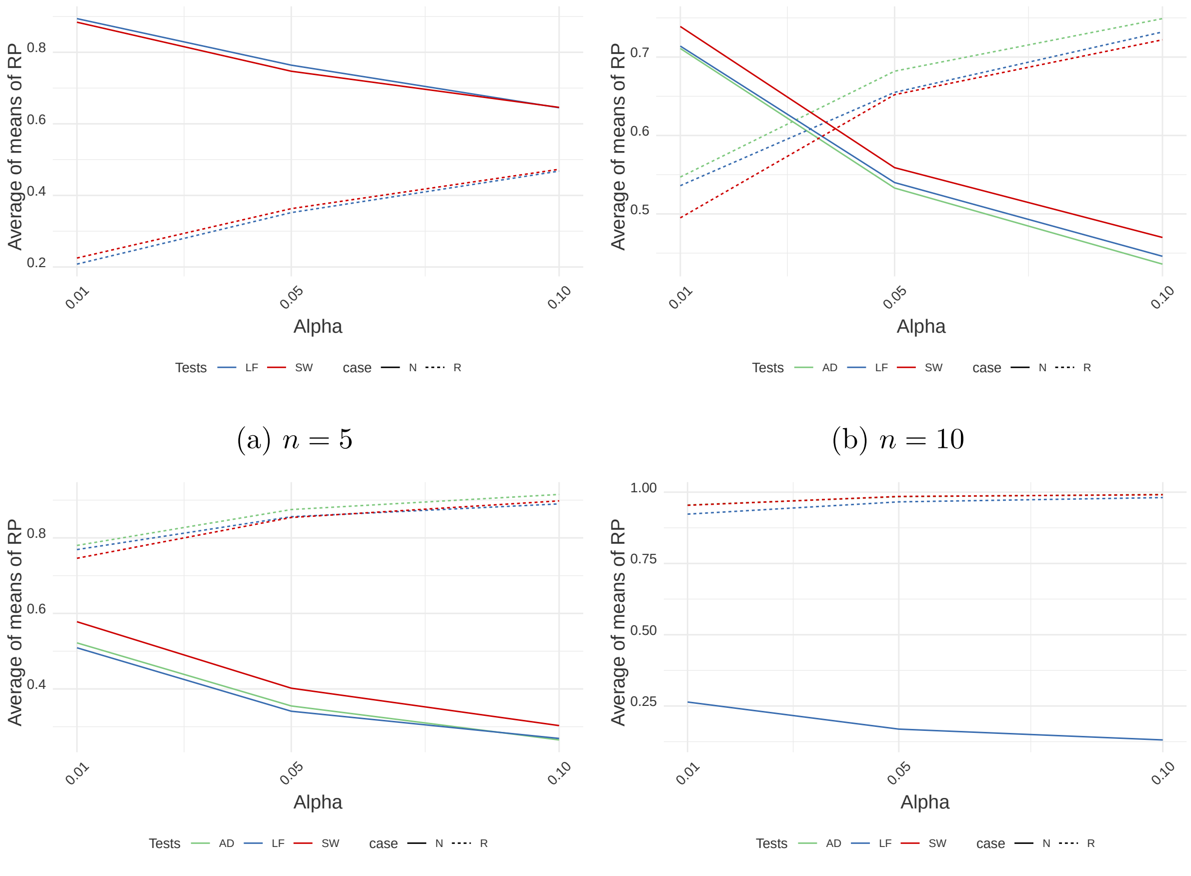


Figure A.4: The relationship between NPI-B-RP and p-value for Shapiro-Wilk and Lilliefors test for data sampled from a mixture of normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, with $n = 50$ and $\alpha = 0.05$

A.2 Reproducibility of Normality tests for different levels of significance

The simulation study discussed in Section 2.3 was repeated under H_1 for the Normality test when data are sampled from $Exp(1)$, using various significance levels: specifically, 0.01, 0.05, and 0.1. From Figure A.5 it can be seen generally that in the area of non-rejection, RP values are high at lower significance levels and decrease as the level of significance increases. On the other hand, within the rejection area, RP values are typically low at small significance levels and increase with higher significance levels. For samples of size 5, the LF test has a slightly higher RP than the SW test in the non-rejection area when $\alpha = 0.01$ and $\alpha = 0.05$. However, at $\alpha = 0.1$ both tests have approximately similar mean of RP values. Conversely, in the rejection area, the SW test has a slightly higher RP than the LF test at significance levels of $\alpha = 0.01$ and $\alpha = 0.05$. Nevertheless, at $\alpha = 0.1$ both tests have approximately the same mean of RP values. The AD test has the highest mean of RP values in the rejection area for sample sizes of 10, and 20. However, for the sample size of 50 AD and SW tests have approximately similar RP values at all levels of significance. In the non-rejection area, the SW test has the highest mean of RP values for sample sizes of 10 and 20 at all values of α. However, for the sample size of 50, there are only RP values for the LF test at all levels of significance.



(a) $n = 5$

(b) $n = 10$

(c) $n = 20$

(d) $n = 50$

Figure A.5: The mean of means RP values for SW, AD and LF tests in both cases rejection (R) and non-rejection(N), sampled from $Exp(1)$ for different levels of α.

Appendix B

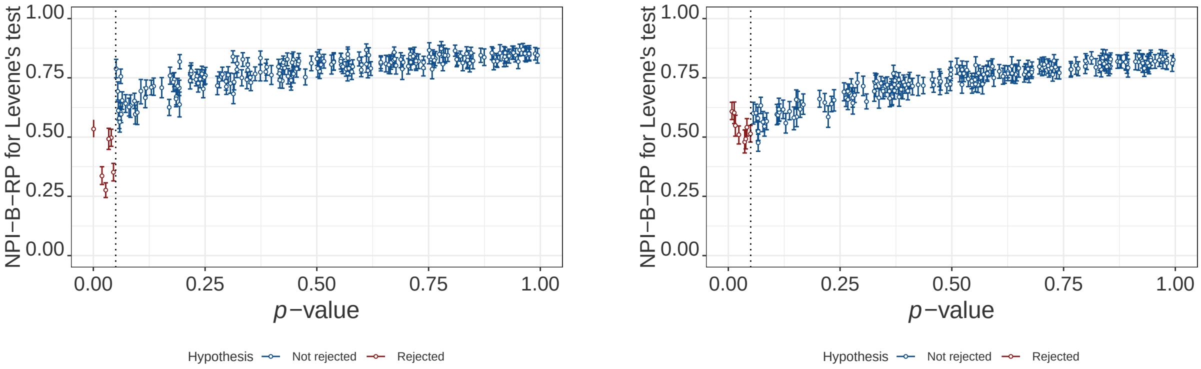
Reproducibility for Equality of Variances Tests

B.1 Levene's test for equality of three variances

This section provides the reproducibility probability (RP) of Levene's test results from the simulation study when comparing the variances of the three original samples.

Figure B.1 shows the relationship between p-values and RP values for Levene's test, under H_0, for the two-sided hypothesis. This relationship shows that RP is low when the p-value is close to the level of significance, and RP tend to be high as the p-value is far away from the level of significance. RP values in the rejection area tend to be low, especially for the sample size of 10. Similarly, under H_1, RP values for Levene's test show the general pattern as presented in Figure B.2. For the sample size of 10, RP has strong variability, especially when the p-value is close to the threshold. For the sample size of 25, there is no noticeable variability in RP and all original samples are located in the rejection area.

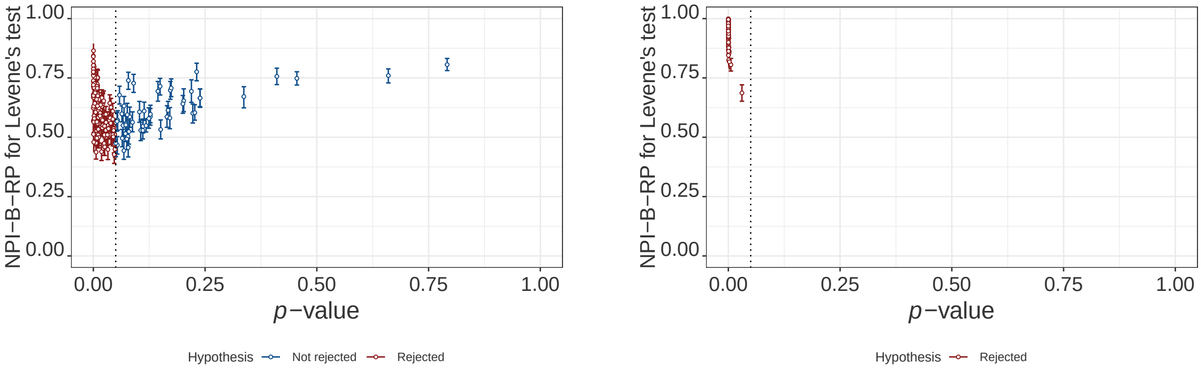
When dealing with non-Normal data and simulating under H_0, the RP values for Levene's test show the general pattern for RP as observed in Figure B.3. There is variability in RP values that are close to the threshold for sample size 10, however, for sample size 25 RP values are more stable.



(a) $n = 10$

(b) $n = 25$

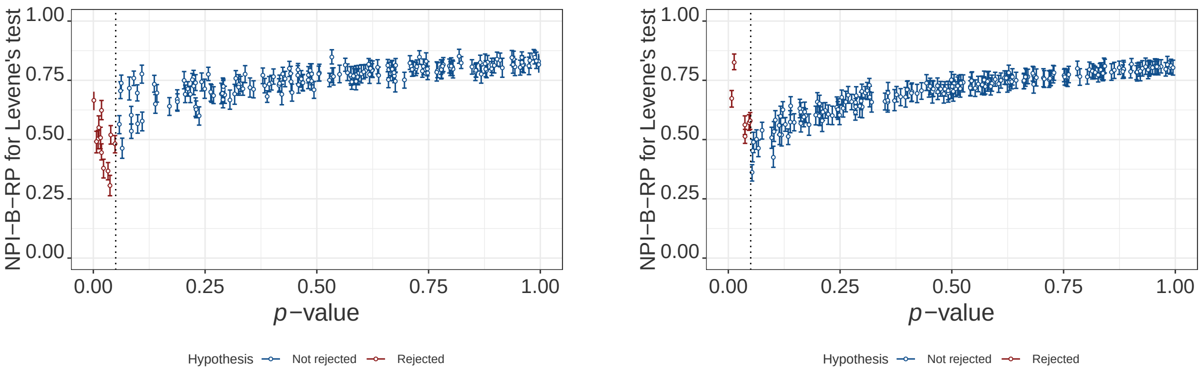
Figure B.1: The relationship between p-values and NPI-B-RP for Levene's test, under H_0, samples sampled from $N(0, 1^2)$, in a two-sided test



(a) $n = 10$

(b) $n = 25$

Figure B.2: The relationship between p-values and NPI-B-RP for Levene's test, under H_1, samples sampled from $N(0, 1^2)$, $N(0, 2^2)$ and $N(0, 4^2)$, in a two-sided test.



(a) $n = 10$

(b) $n = 25$

Figure B.3: The relationship between p-values and NPI-B-RP for Levene's test, samples sampled from $Exp(1)$ under H_0, in a two-sided test

Appendix C

Reproducibility for the One-Sample Location Tests with and without Preliminary Test

C.1 Flowcharts for assessing RP

This section provides flowcharts showing how to evaluate the reproducibility for an example for *Cases A, B and C*.

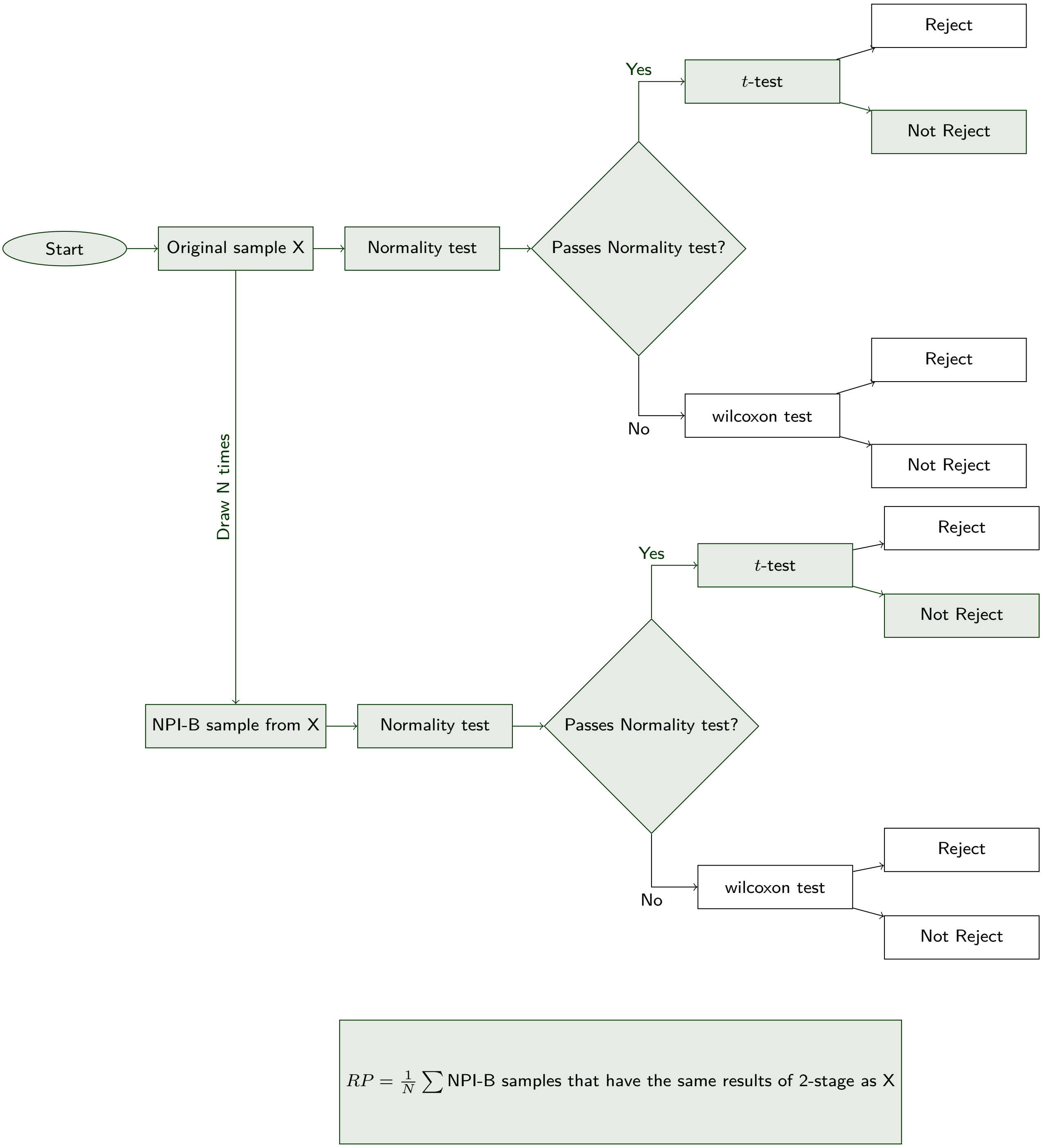


Figure C.1: An illustrative example of the reproducibility assessment for the two-stage test of one-sample location test for Case A.

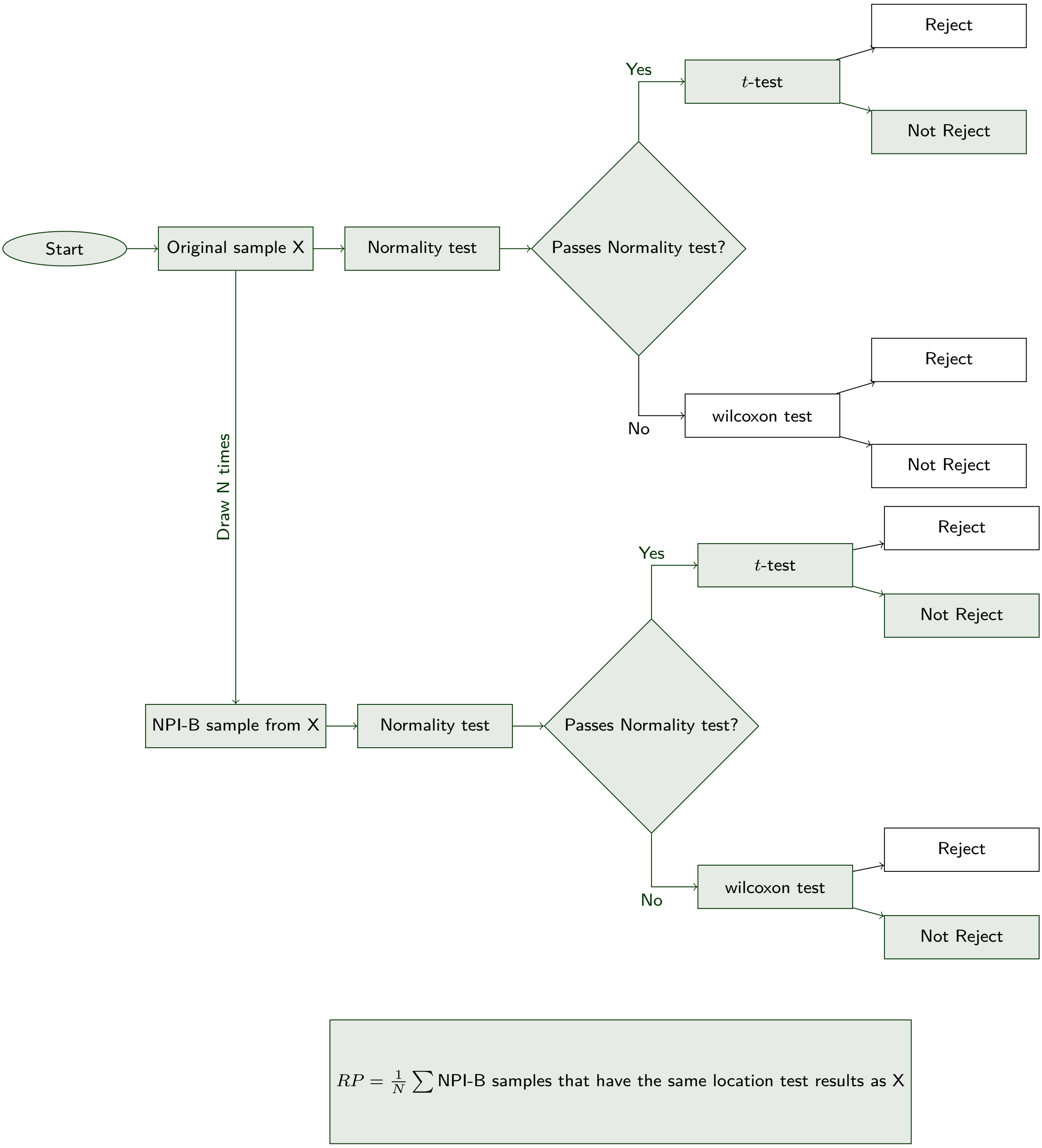


Figure C.2: An illustrative example of the reproducibility assessment for the two-stage test of one-sample location test for Case B.

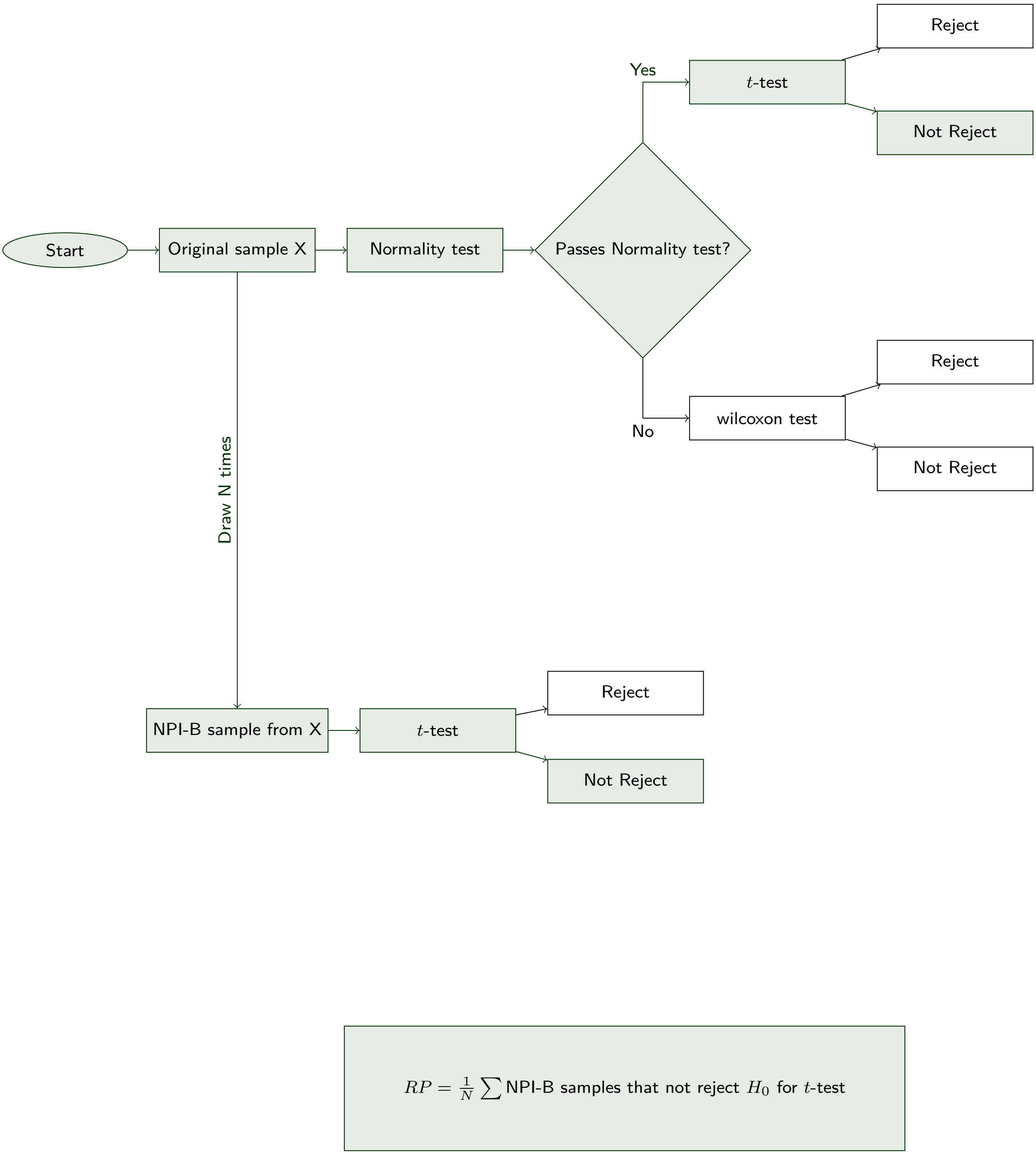


Figure C.3: An illustrative example of the reproducibility assessment for the two-stage test of one-sample location test for Case C.

Appendix D

Reproducibility for the two-Sample Location Tests with and without Preliminary Tests

D.1 Extended simulation examples

This section shows the results for RP values for two-sample location tests with and without preliminary tests when data generated from a mixture of Normal distributions $X \sim 0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$. The results are presented in Figures D.1- D.7. The RP values are approximately similar to those when data are generated from an unimodal distribution shown in the main examples discussed in Chapter 5.

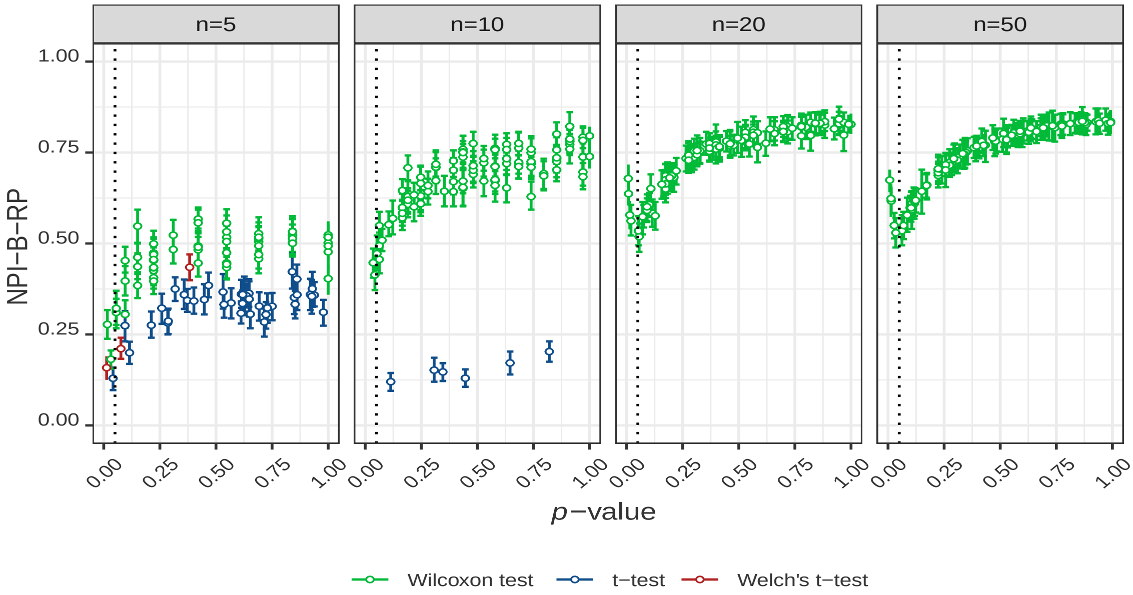


Figure D.1: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case A*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

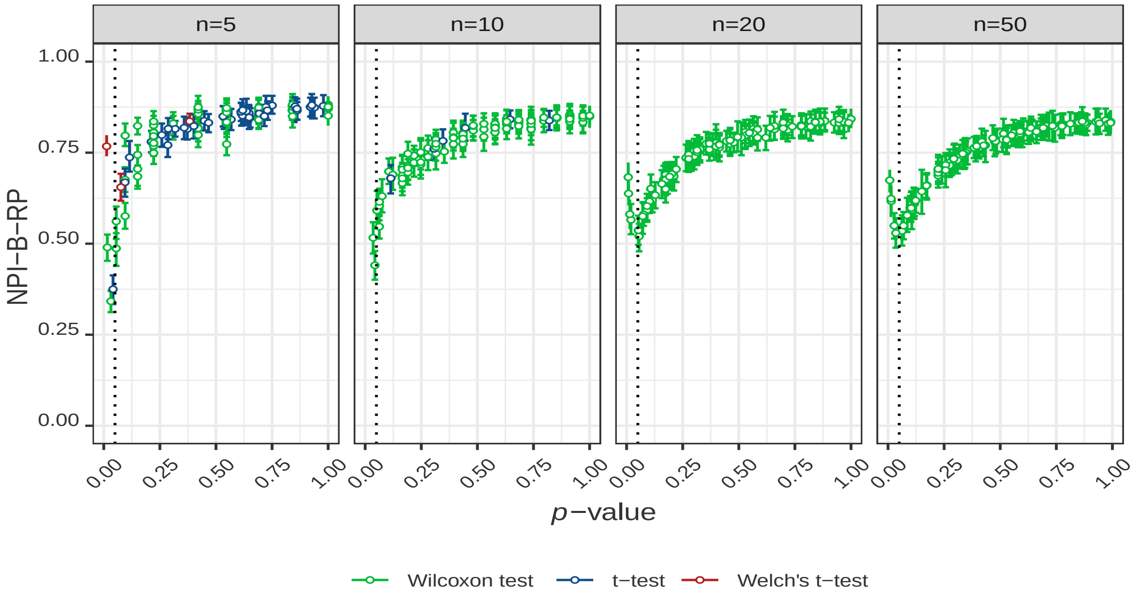


Figure D.2: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case B*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

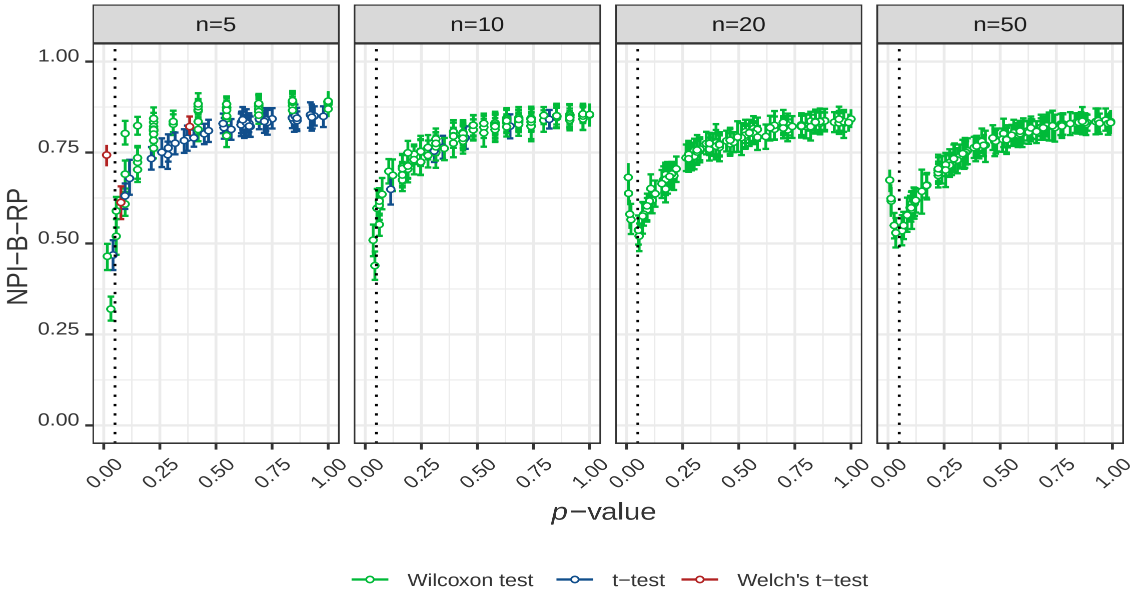


Figure D.3: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case C*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

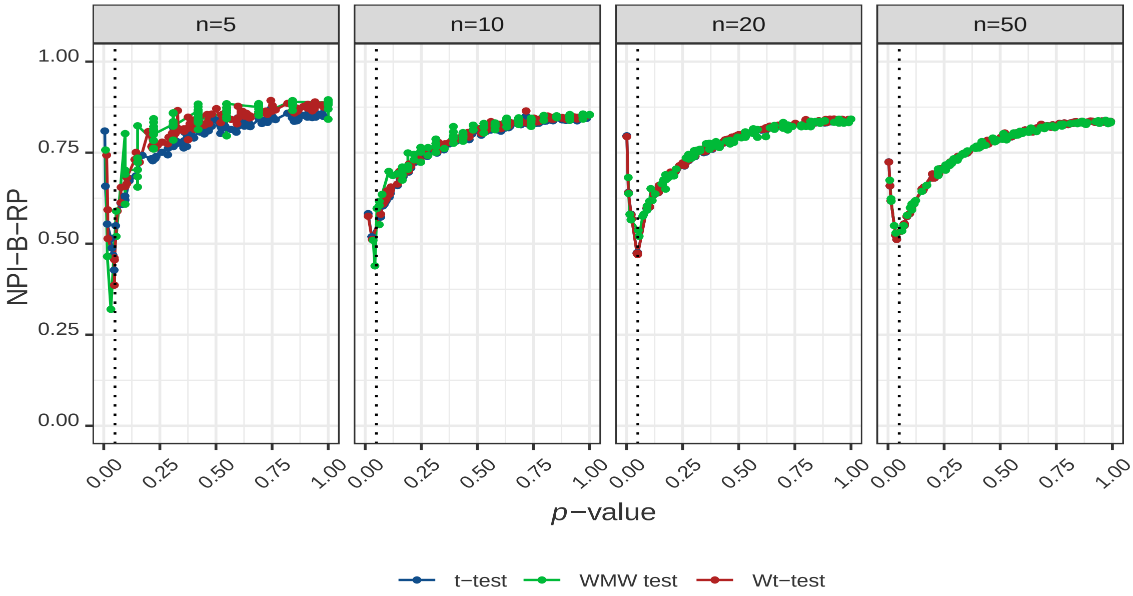


Figure D.4: The means of RP values against the p-values for the location tests without preliminary tests, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

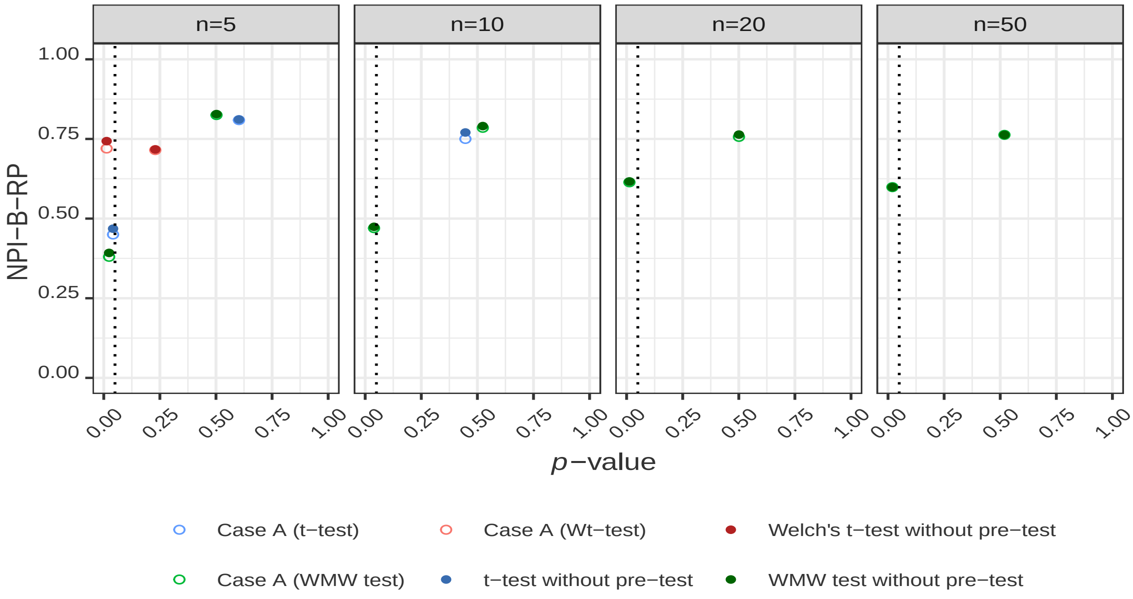


Figure D.5: Comparing RP values for location tests with and without preliminary test (*Case A*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

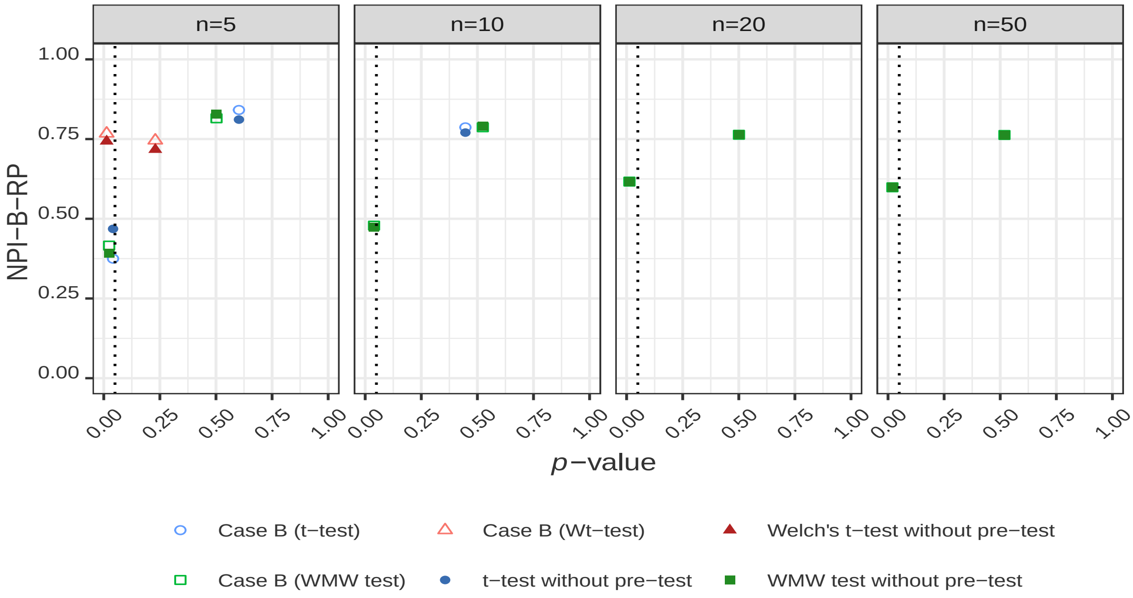


Figure D.6: Comparing RP values for location tests with and without preliminary test (*Case B*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

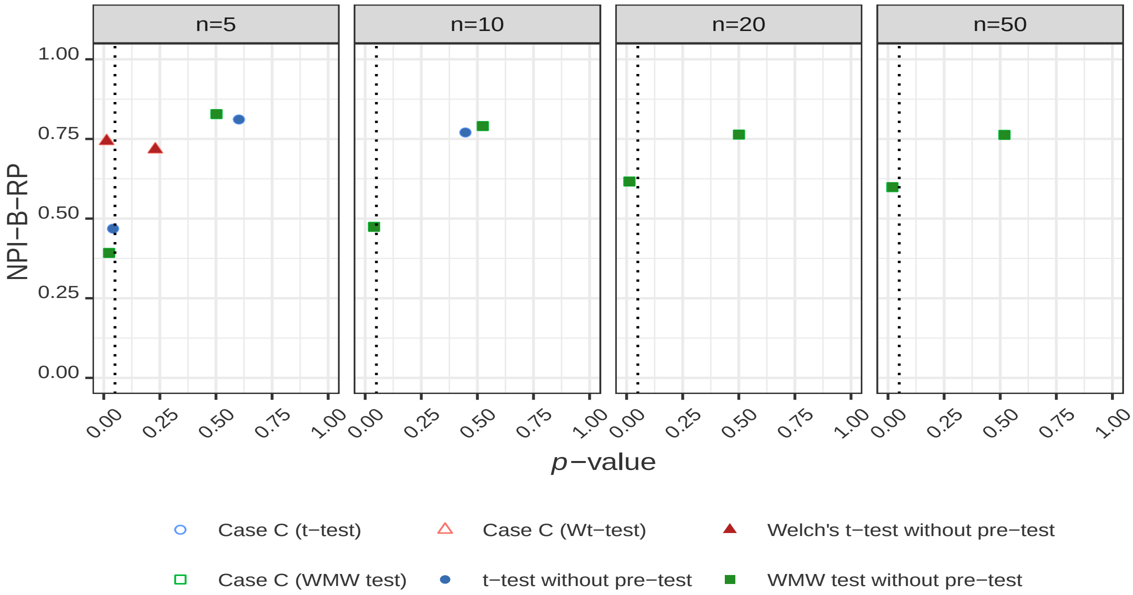


Figure D.7: Comparing RP values for location tests with and without preliminary test (*Case C*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$.

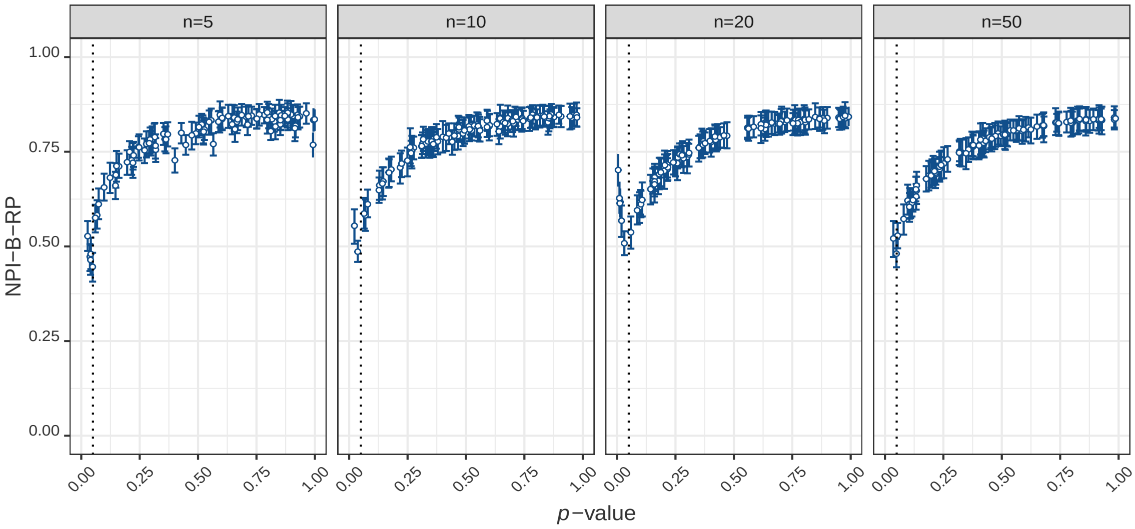


Figure D.8: The min, mean, and max of RP values for the t-test against p-values of the t-test, when original samples are sampled from $N(0, 1)$.

D.2 The results of the reproducibility for the location tests without preliminary tests

This section shows the results of the reproducibility for the two-sample location tests (the two-sample t-test, Welch's t-test, and WMW test) conducted without preliminary tests.

Figures D.8, D.9, and D.10 show the results of simulations for RP for the location tests without preliminary tests. When both original samples are drawn from Normal distributions having identical mean values and equality variances $N(0, 1)$ (where $\mu_X = \mu_Y = 0$, and $\sigma_X = \sigma_Y = 1$).

Figures D.11, D.12, and D.13 present simulation results for the RP of location tests without preliminary tests. The simulation was conducted with original samples drawn from Normal distributions $N(0, 1)$ and $N(1, 2^2)$.

Figures D.14, D.15, and D.16 present simulation results for the RP of location tests without preliminary tests. The simulations were conducted with original samples drawn from non-Normal distributions with the same means and variances $LN(0, 1)$ (where $\mu_X = \mu_Y \approx 1.648$ and $\sigma_X = \sigma_Y \approx 2.944$).

Figures D.17, D.18, and D.19 present simulation results for the RP of location tests

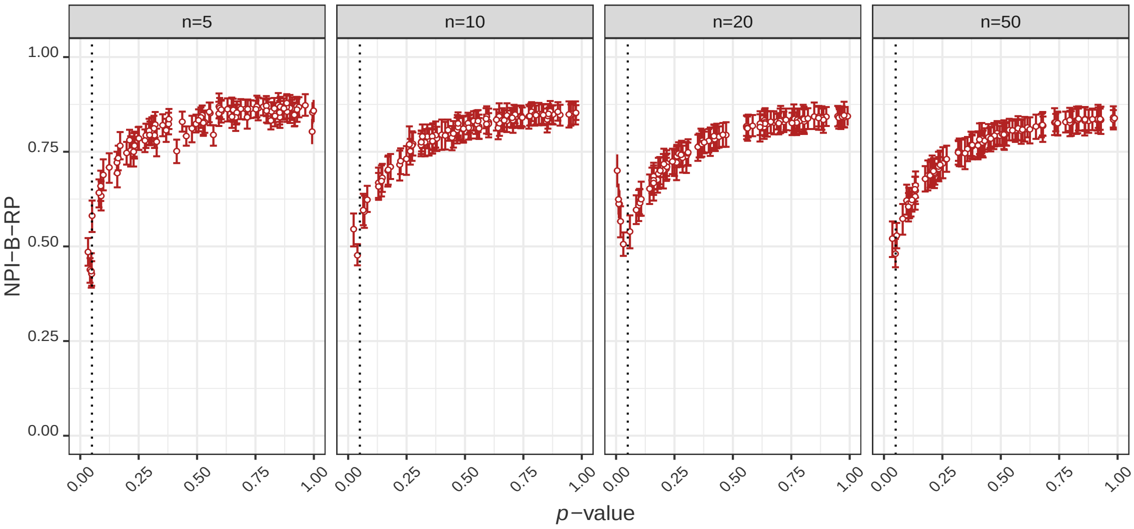


Figure D.9: The min, mean, and max of RP values for the Welch's t-test against p-values of the Welch's t-test, when original samples are sampled from $N(0,1)$.

t-test, Welch's t-test, and WMW test without preliminary tests, respectively. The simulations were conducted with original samples drawn from non-Normal distributions with different means and variances $LN(0,1)$ and $LN(1,0.5)$ (where $\mu_X \approx 1.648, \mu_Y \approx 3.0689$ and $\sigma_X \approx 2.944, \sigma_Y \approx 5.575$).

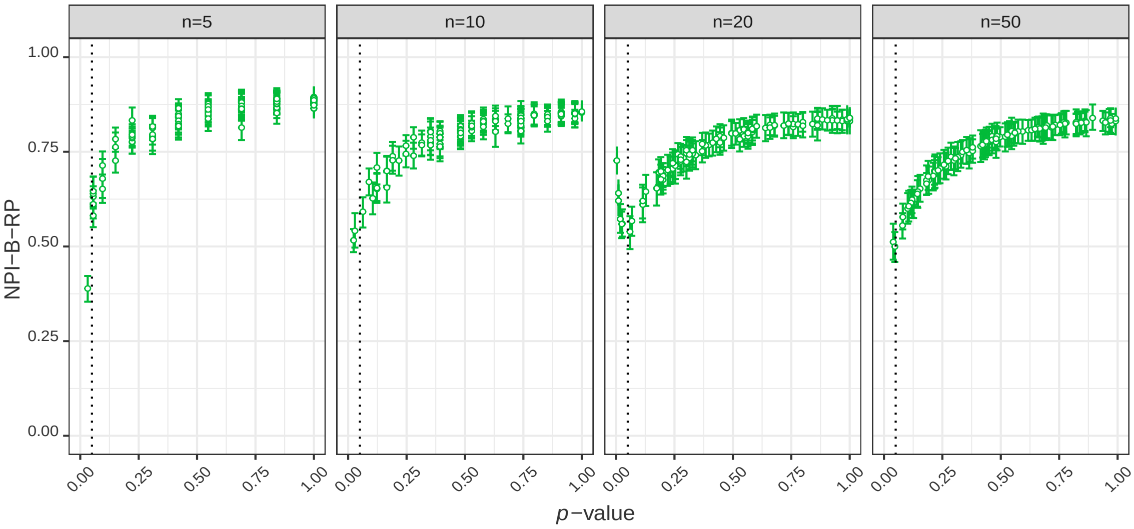


Figure D.10: The min, mean, and max of RP values for the WMW test against p-values of the WMW test, when original samples are sampled from $N(0, 1)$.

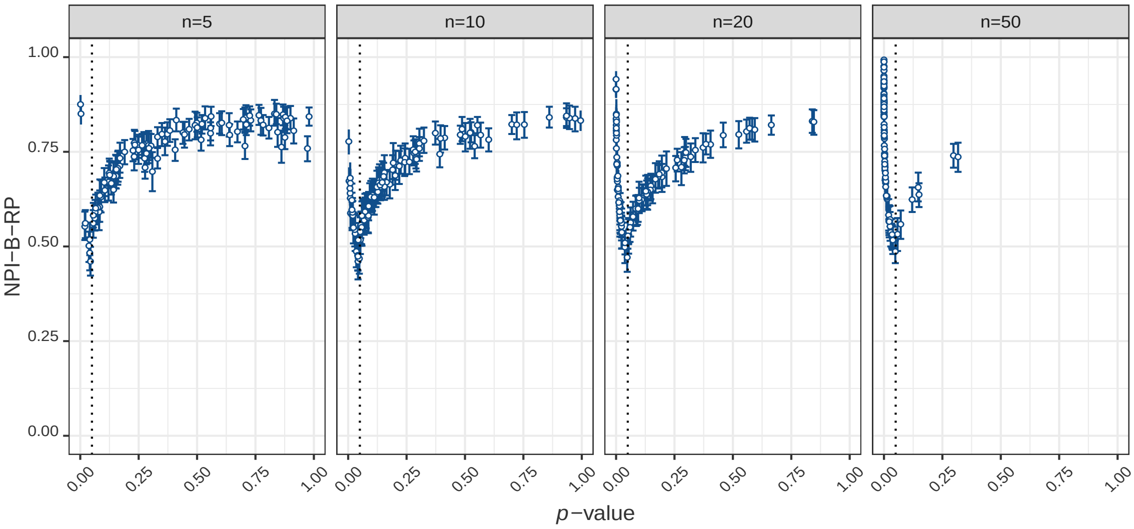


Figure D.11: The min, mean, and max of RP values for the t-test against p-values of the t-test, when original samples are sampled from $N(0, 1)$ and $N(1, 2^2)$.

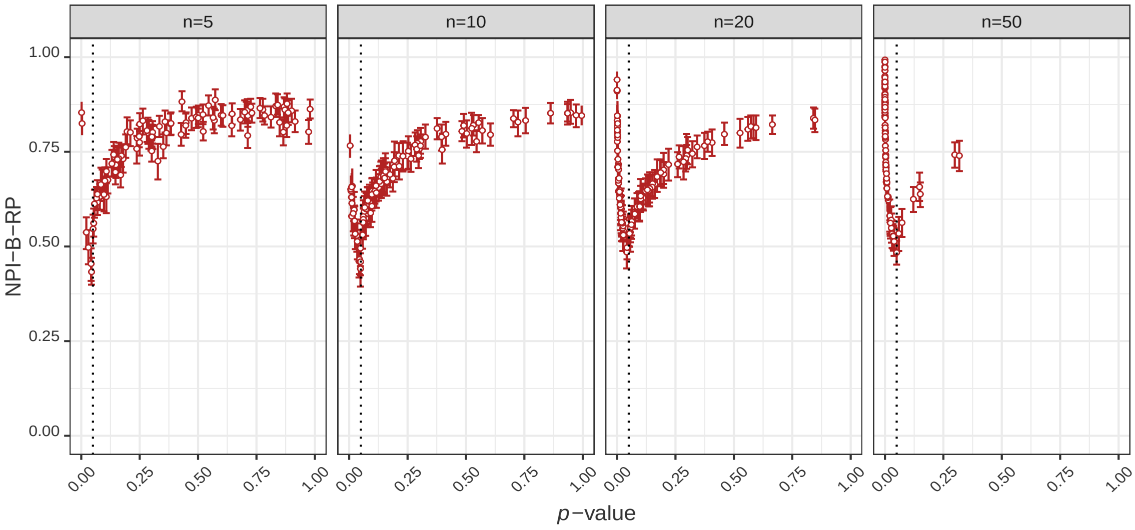


Figure D.12: The min, mean, and max of RP values for the Welch's t-test against p-values of the Welch's t-test, when original samples are sampled from $N(0,1)$ and $N(1,2^2)$.

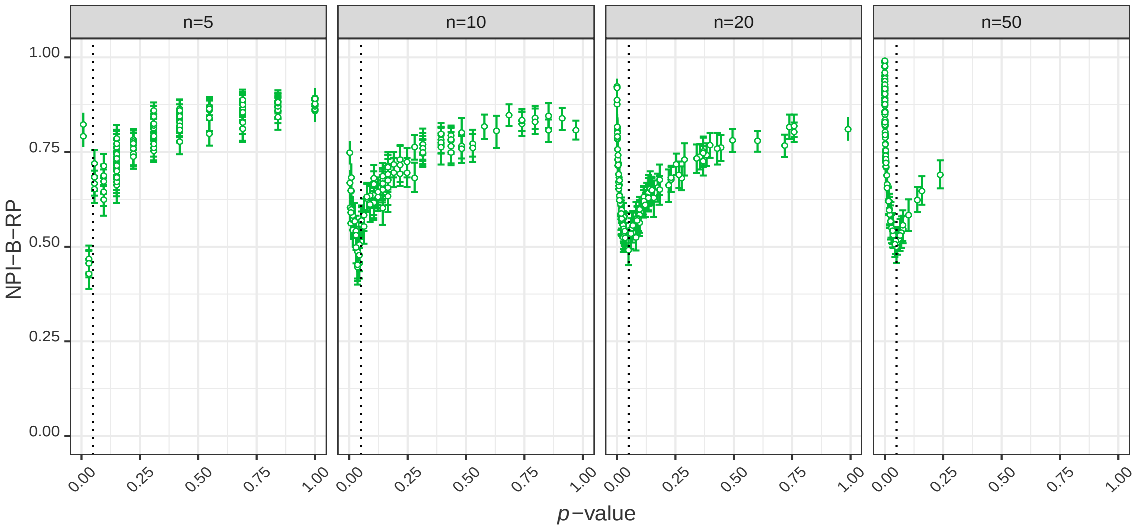


Figure D.13: The min, mean, and max of RP values for the WMW test against p-values of the WMW test, when original samples are sampled from $N(0,1)$ and $N(1,2^2)$.

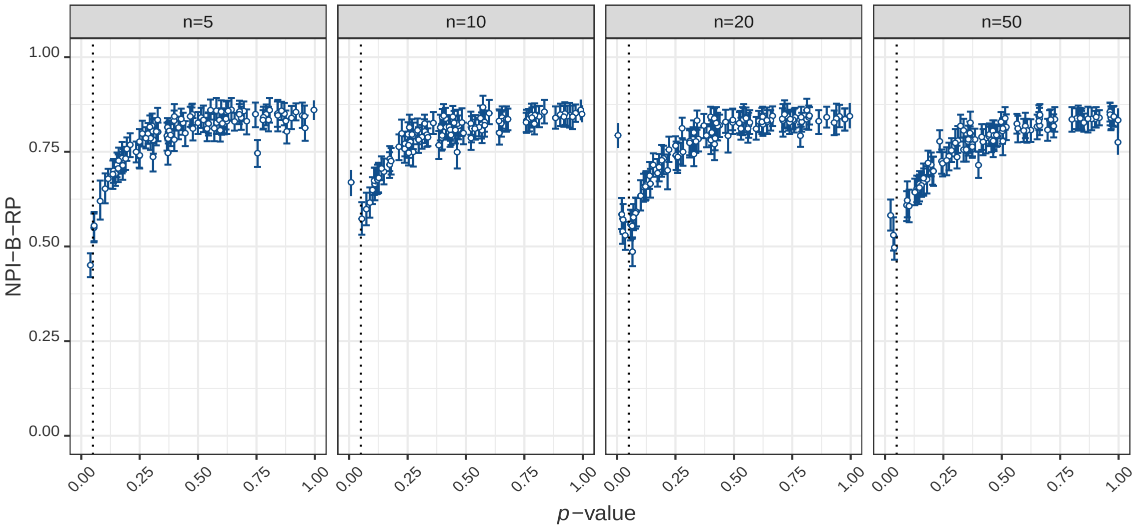


Figure D.14: The min, mean, and max of RP values for the t-test against p-values of the t-test, when original samples are sampled from $LN(0, 1)$.

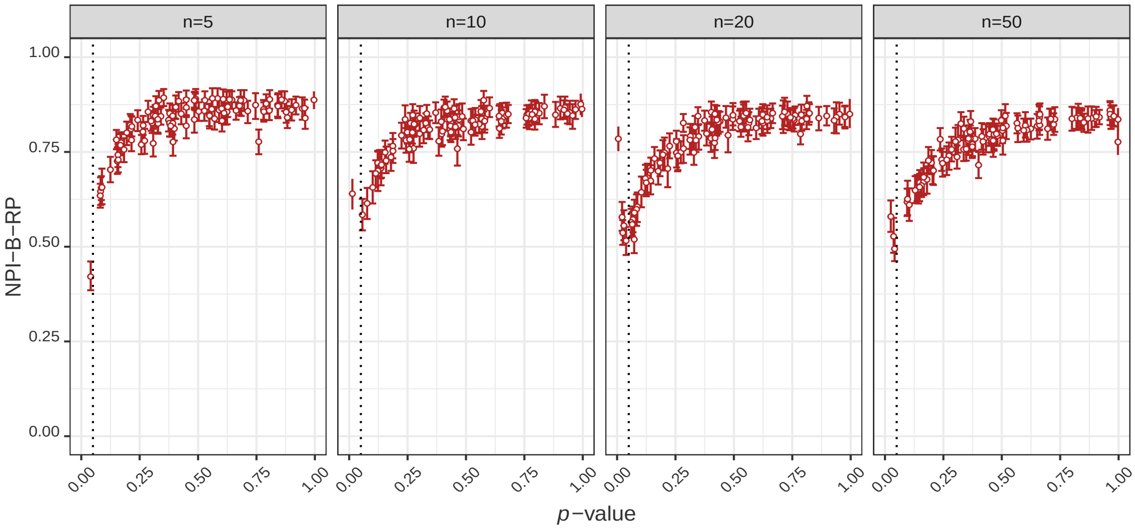


Figure D.15: The min, mean, and max of RP values for the Welch's t-test against p-values of the Welch's t-test, when original samples are sampled from $LN(0, 1)$.

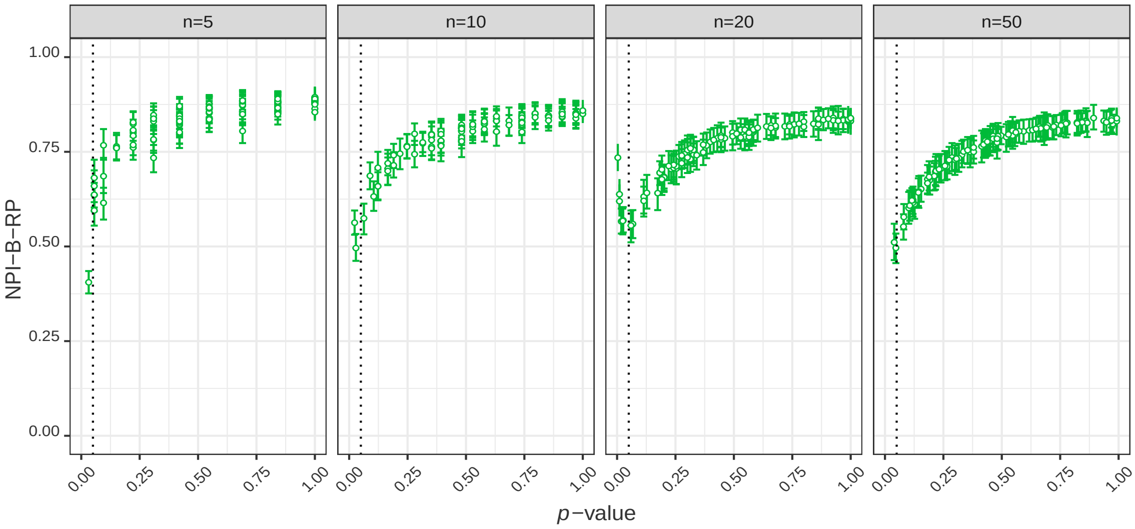


Figure D.16: The min, mean, and max of RP values for the WMW test against p-values of the WMW test, when original samples are sampled from $LN(0, 1)$.

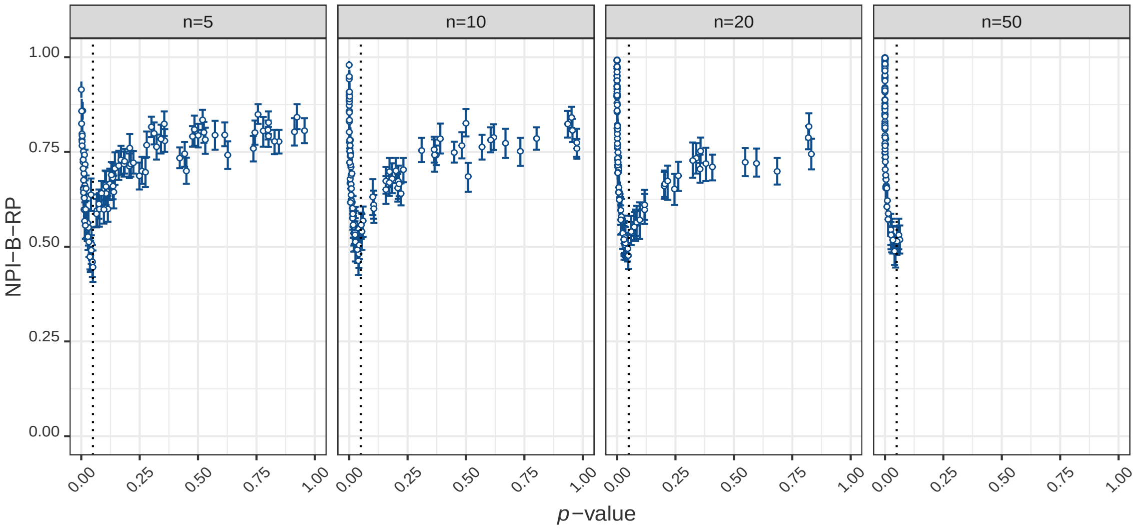


Figure D.17: The min, mean, and max of RP values for the t-test against p-values of the t-test, when original samples are sampled from $LN(0, 1)$ and $LN(1, 0.5)$.

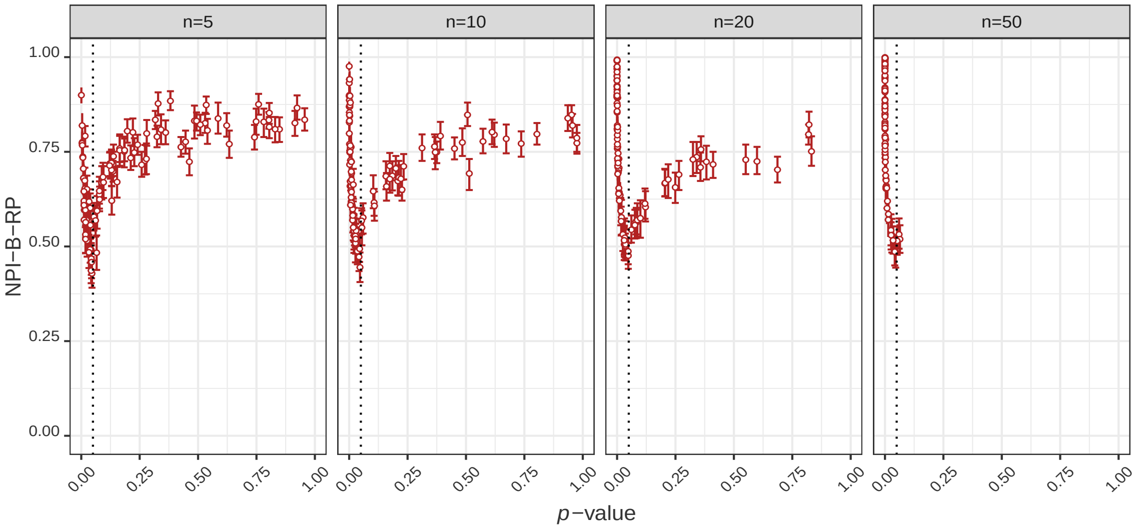


Figure D.18: The min, mean, and max of RP values for the Welch's t-test against p-values the Welch's t-test, when original samples are sampled from $LN(0,1)$ and $LN(1,0.5)$.

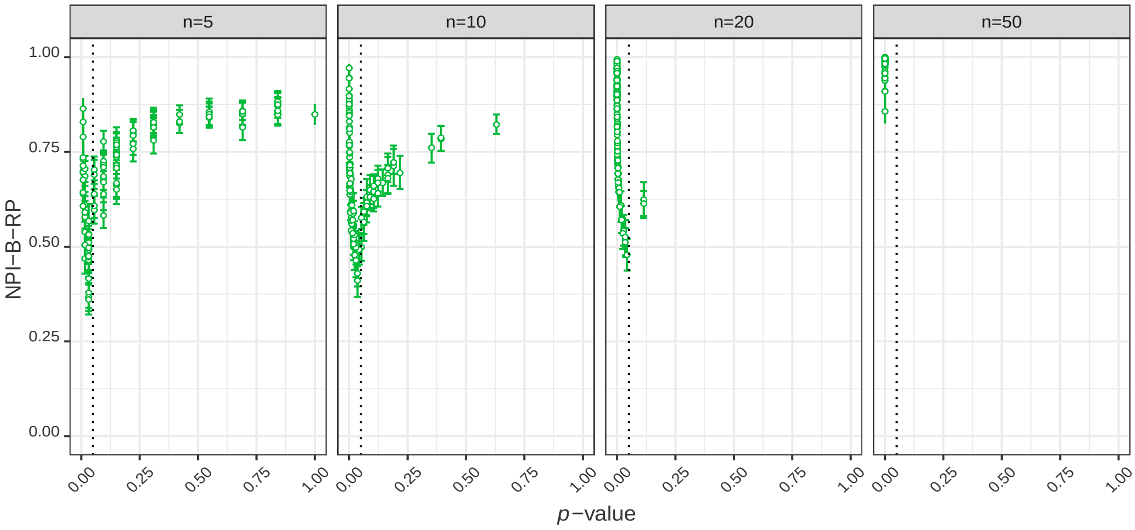


Figure D.19: The min, mean, and max of RP values for the WMW test against p-values of the WMW test, when original samples are sampled from $LN(0,1)$ and $LN(1,0.5)$.

Appendix E

Reproducibility for the Multiple-Sample Location Tests with and without Preliminary Tests

E.1 The impact of the preliminary test of Normality on RP of location tests

E.1.1 The impact of the preliminary test of Normality on RP of location tests for *Case A*

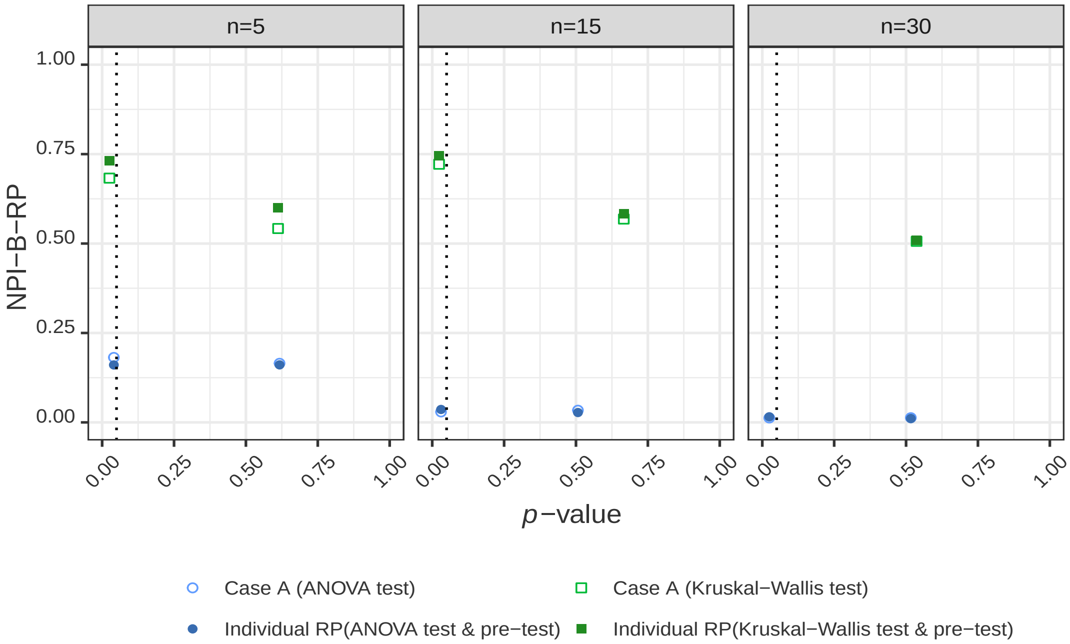


Figure E.1: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $N(0,1)$, $M = 5$.

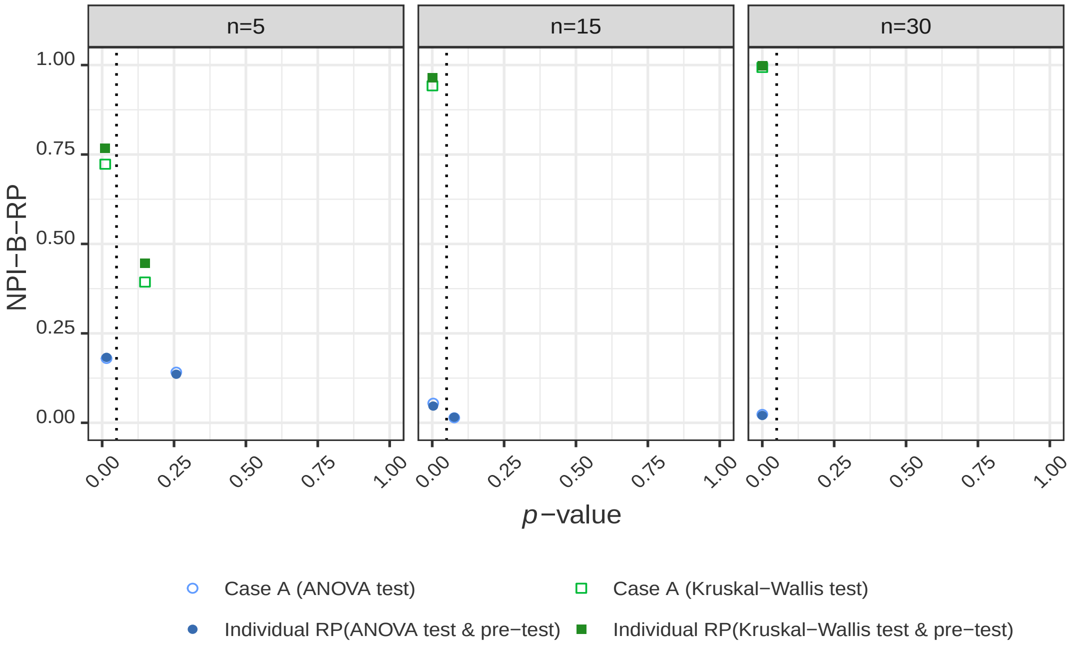


Figure E.2: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $N(1,1)$, $N(0,1)$, $N(2,1)$, $N(2,2)$, and $N(0,2)$, $M = 5$.

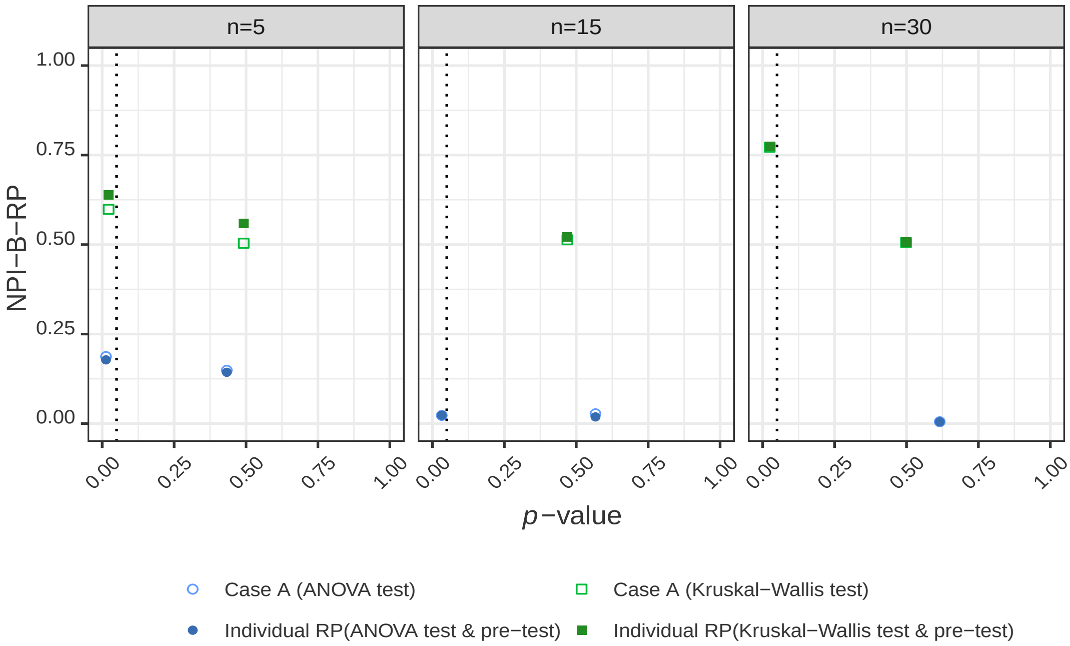


Figure E.3: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $t(4)$, $M = 5$.

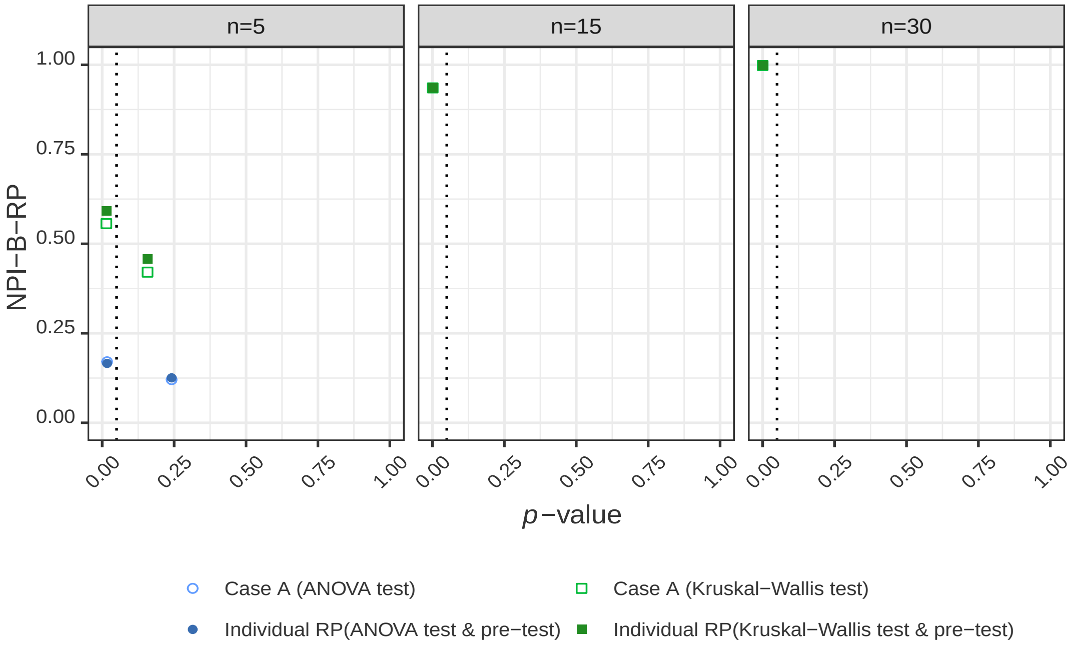


Figure E.4: Comparison of the mean of RP for location tests with and without preliminary test for *Case A* against the mean of their p-values, samples from $N(0,1)$, $N(1,1)$, $LN(1,2)$,$Ca(1,1)$, and $t(4)$,$M = 5$.

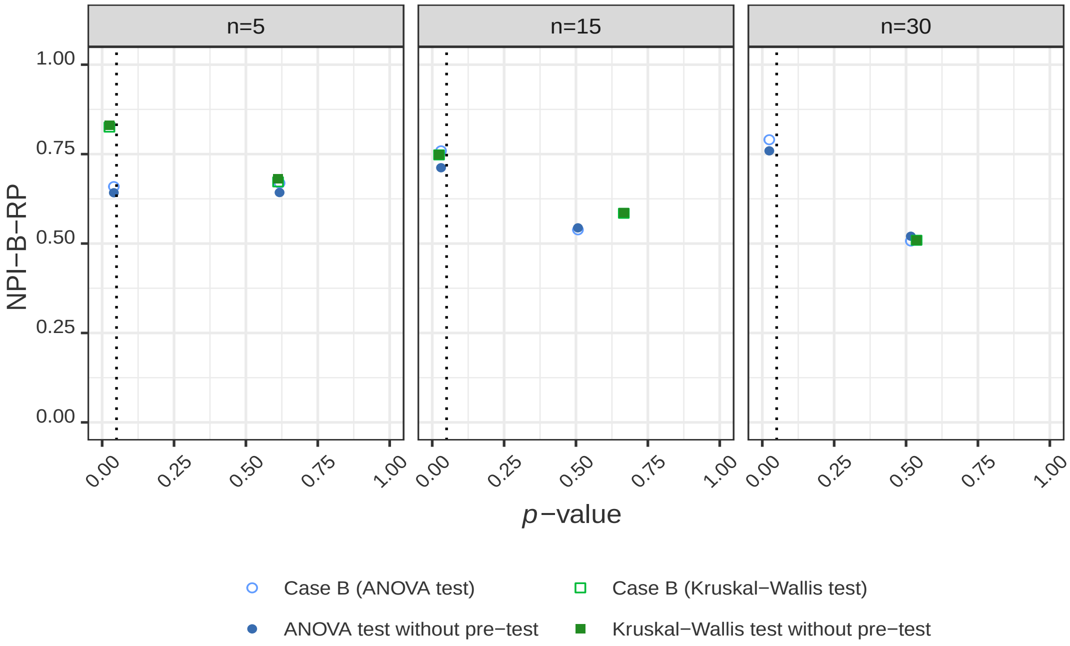


Figure E.5: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $N(0,1)$, $M = 5$.

E.1.2 The impact of the preliminary test on RP of location tests for *Case B*

The results of comparing the reproducibility of the outcome of the location tests (*Case B*) with the reproducibility of location tests without the preliminary test of Normality for 5 groups are presented.

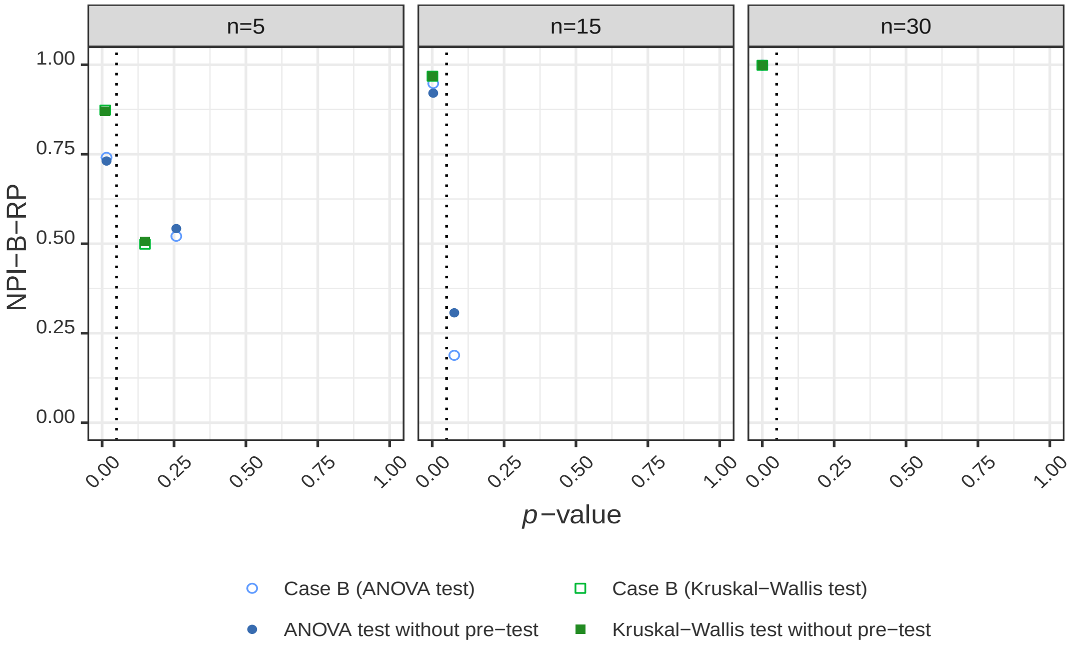


Figure E.6: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $N(1,1)$, $N(0,1)$, $N(2,1)$,$N(2,2)$, and $N(0,2)$, $M = 5$.

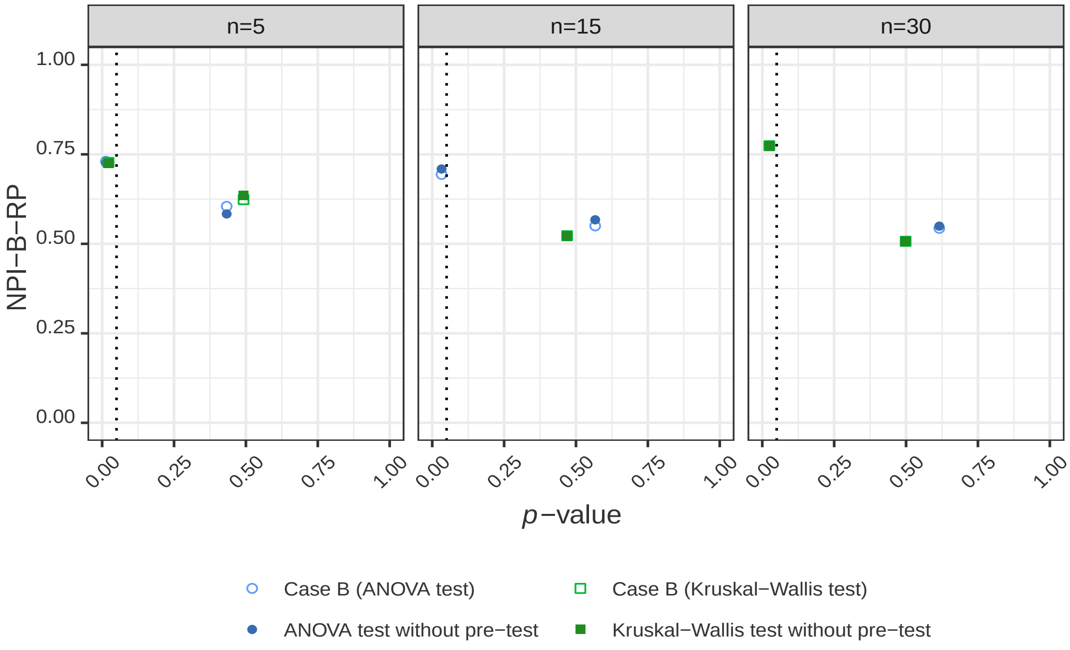


Figure E.7: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $t(4)$, $M = 5$.

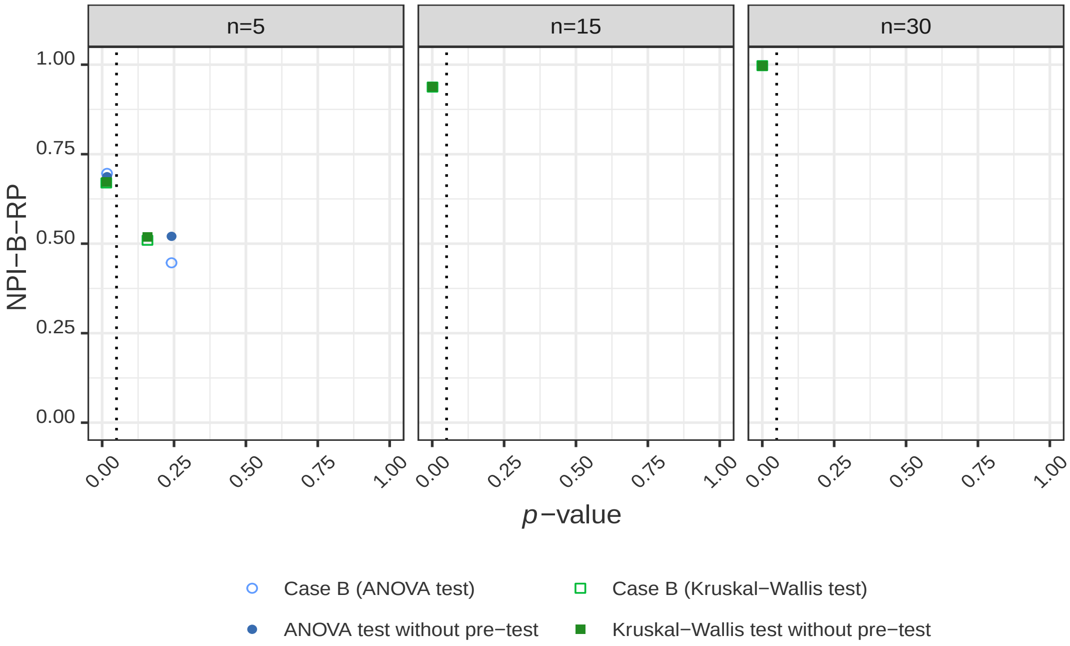


Figure E.8: Comparison of the mean of RP for location tests with and without preliminary test (*Case B*) against the mean of their p-values, samples from $N(0,1)$, $N(1,1)$, $LN(1,2)$, $Ca(1,1)$, and $t(4)$, $M = 5$.

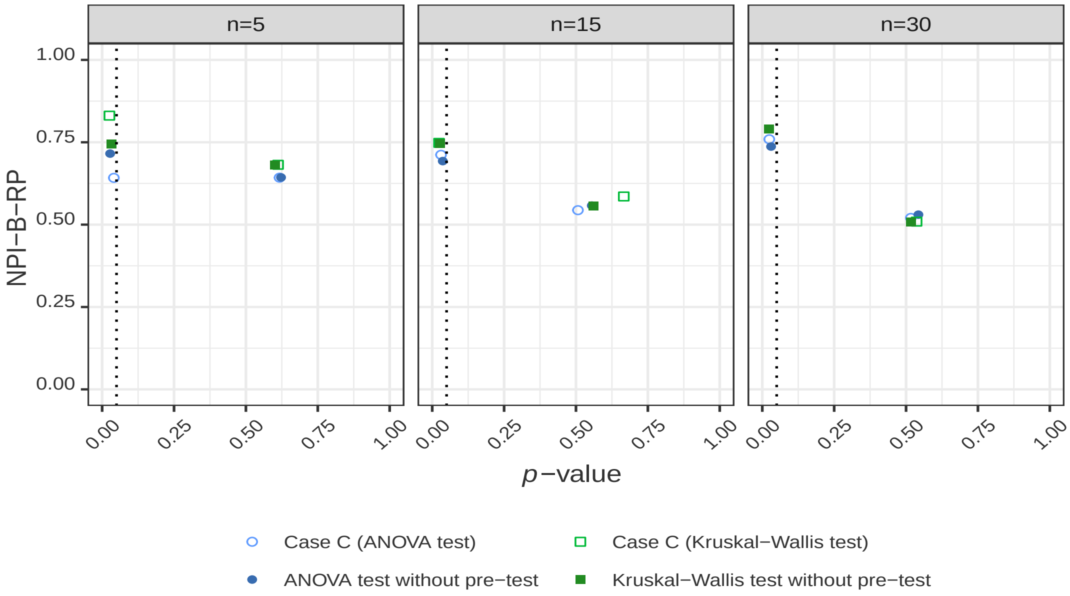


Figure E.9: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $N(0, 1)$, $M = 5$.

E.1.3 The impact of the preliminary test on RP for location tests for *Case C*

The comparison results for the reproducibility for *Case C* and reproducibility for location tests without preliminary tests are shown for 5 groups.

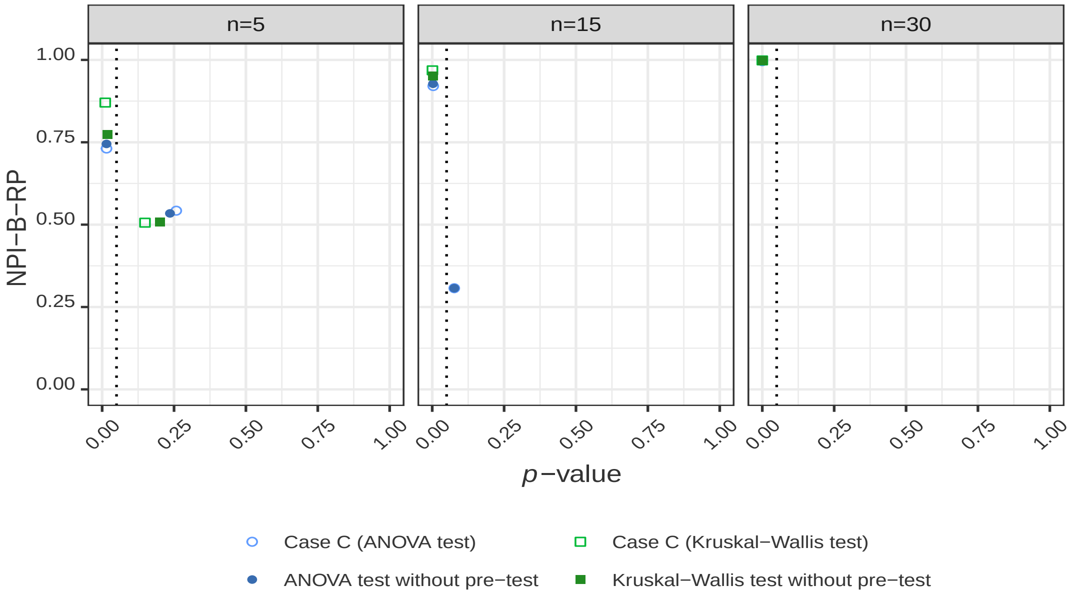


Figure E.10: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $N(1,1)$, $N(0,1)$, $N(2,1)$,$N(2,2)$, and $N(0,2)$, $M = 5$.

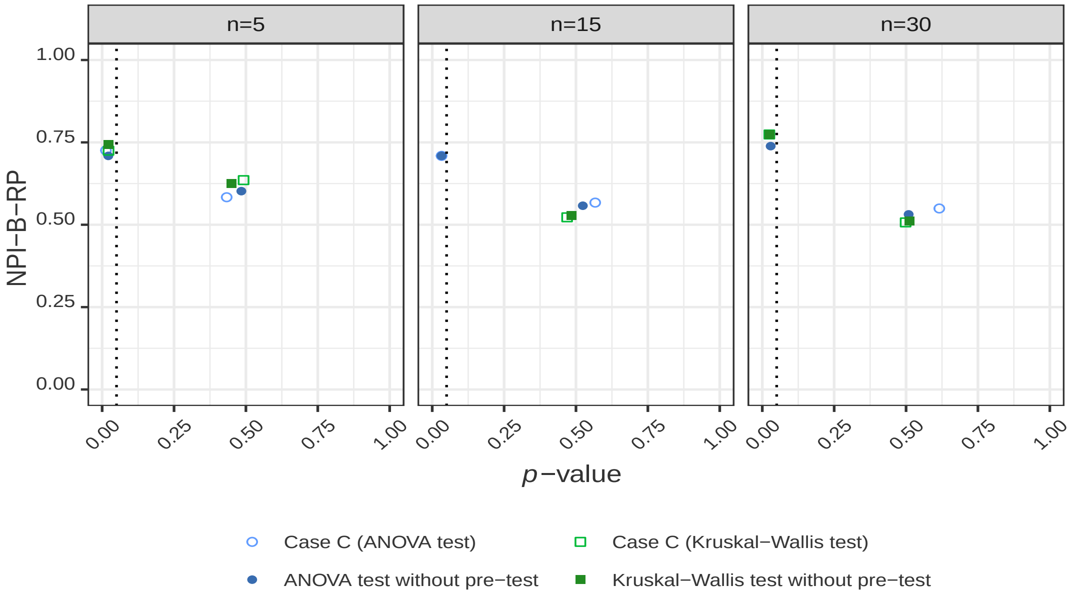


Figure E.11: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $t(4)$, $M = 5$.

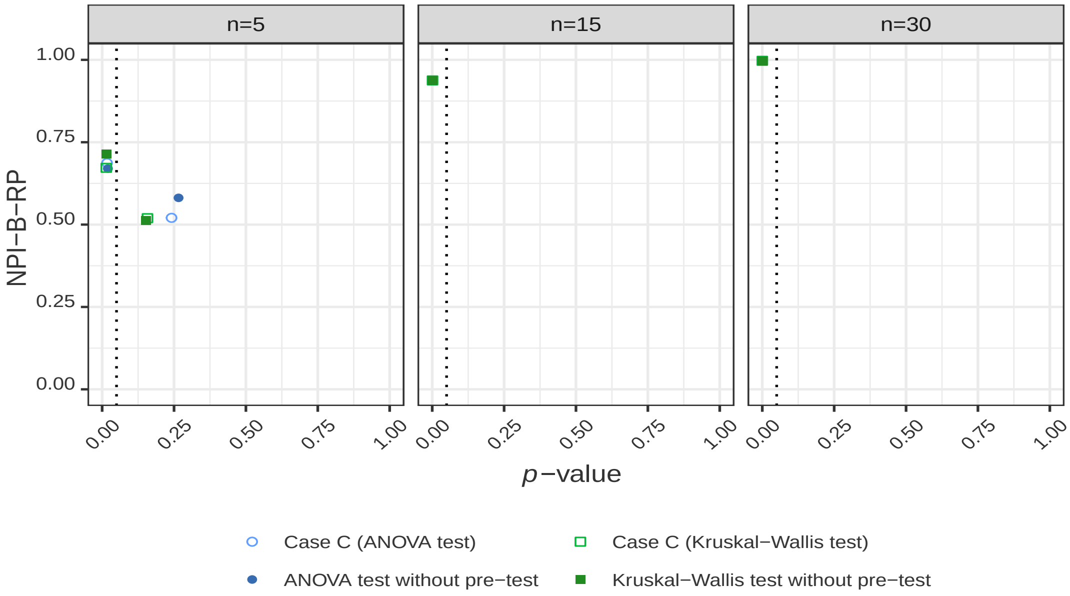


Figure E.12: Comparison of the mean of RP for location tests with and without preliminary test (*Case C*) against the mean of their p-values, samples from $N(0,1)$, $N(1,1)$, $LN(1,2)$, $Ca(1,1)$, and $t(4)$, $M = 5$.

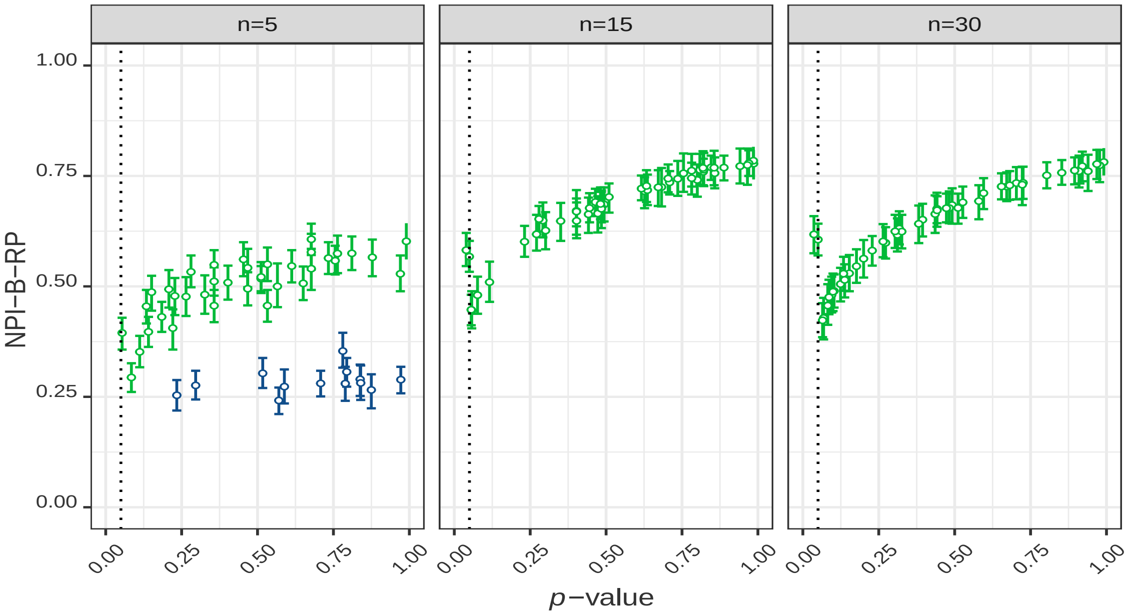


Figure E.13: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case A*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

E.2 Extended simulation examples

This section contains detailed results from the additional simulations involving data generated from a mixture of Normal distributions $X \sim 0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, for $M = 3$. The results are presented in Figures E.13 - E.18. The RP values are approximately similar to those when data are generated from an unimodal distribution, as shown in the main examples discussed in Chapter 6.

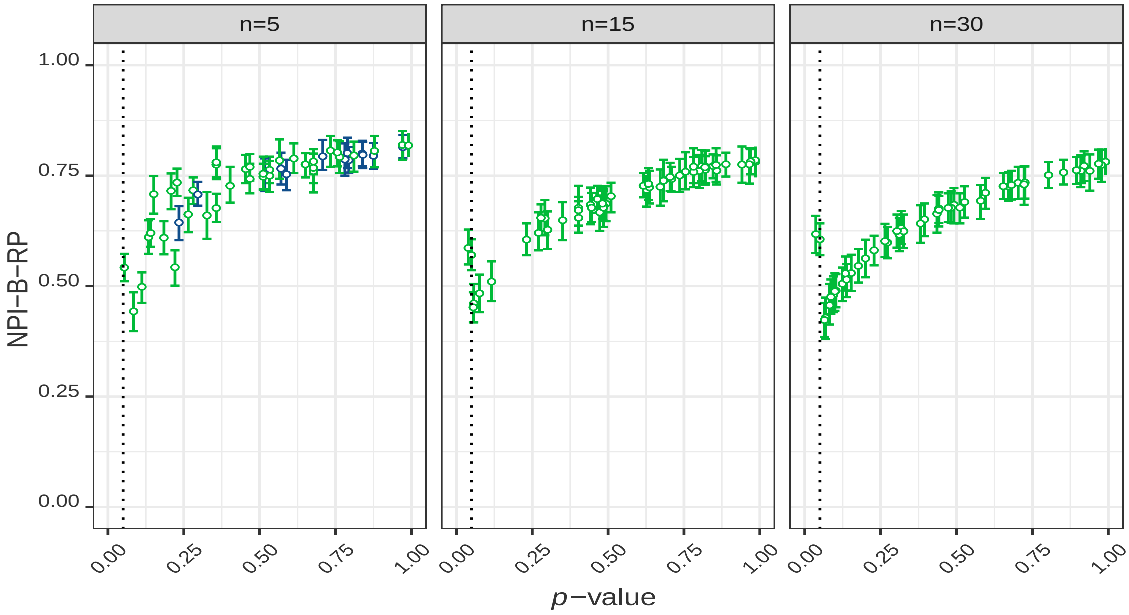


Figure E.14: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case B*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

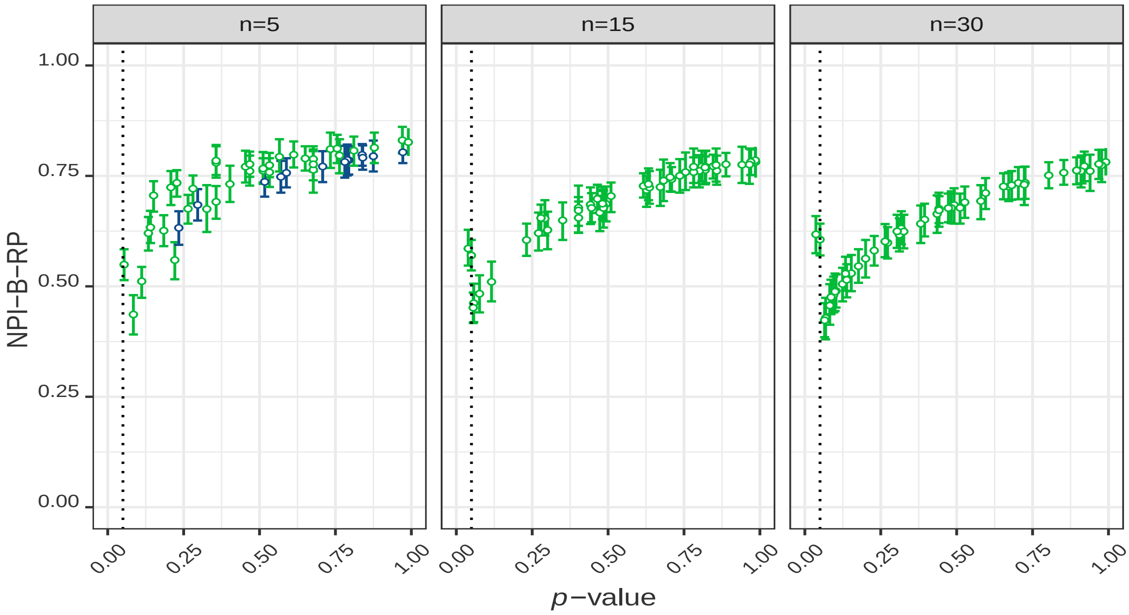


Figure E.15: The min, mean, and max of RP values against the p-values for the two-stage procedure for *Case C*, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

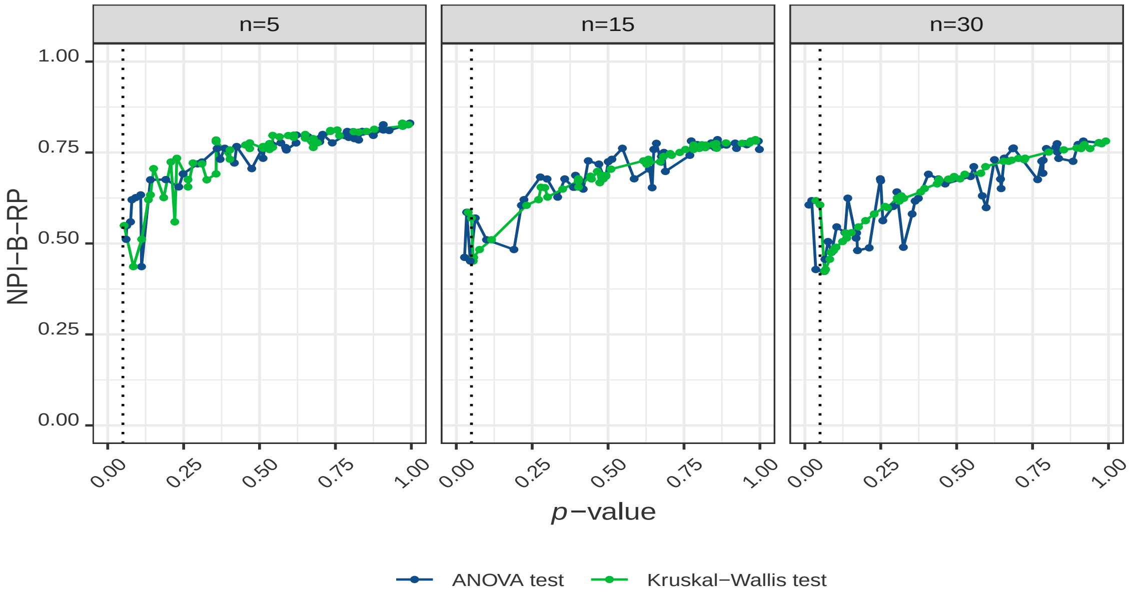


Figure E.16: The means of RP values against the p-values for the location tests without preliminary tests, the original samples are drawn the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

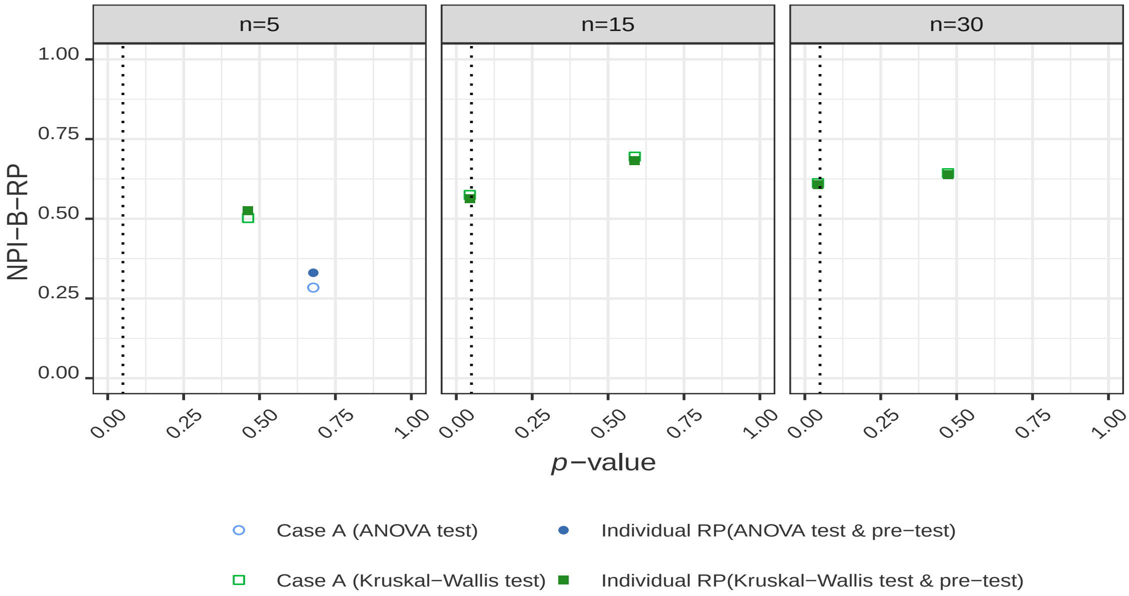


Figure E.17: Comparing RP values for location tests with and without preliminary test (*Case A*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

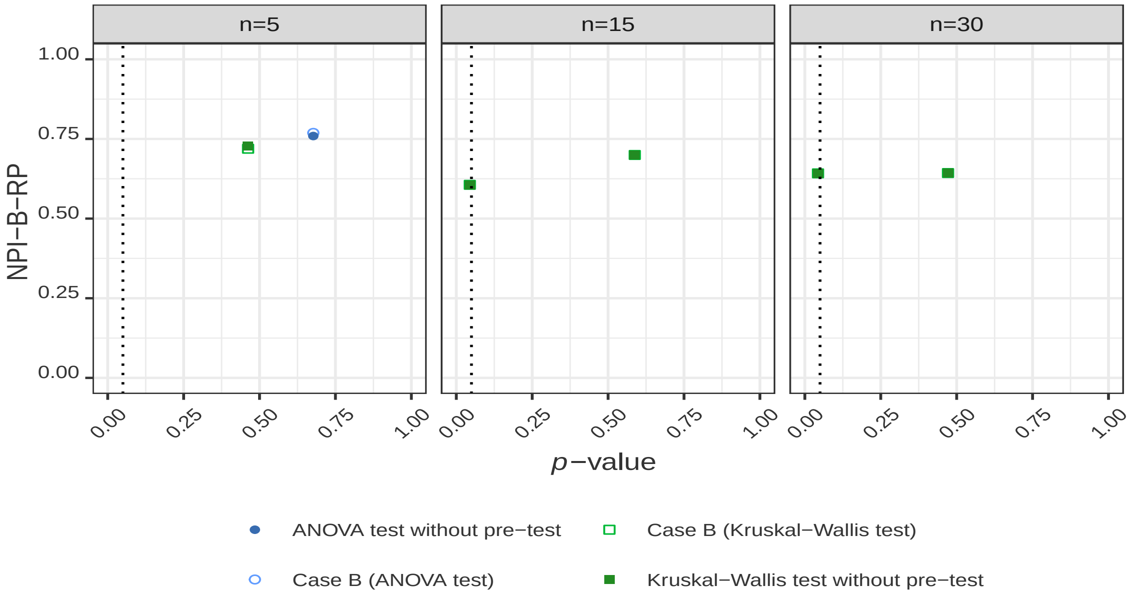


Figure E.18: Comparing RP values for location tests with and without preliminary test (*Case B*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

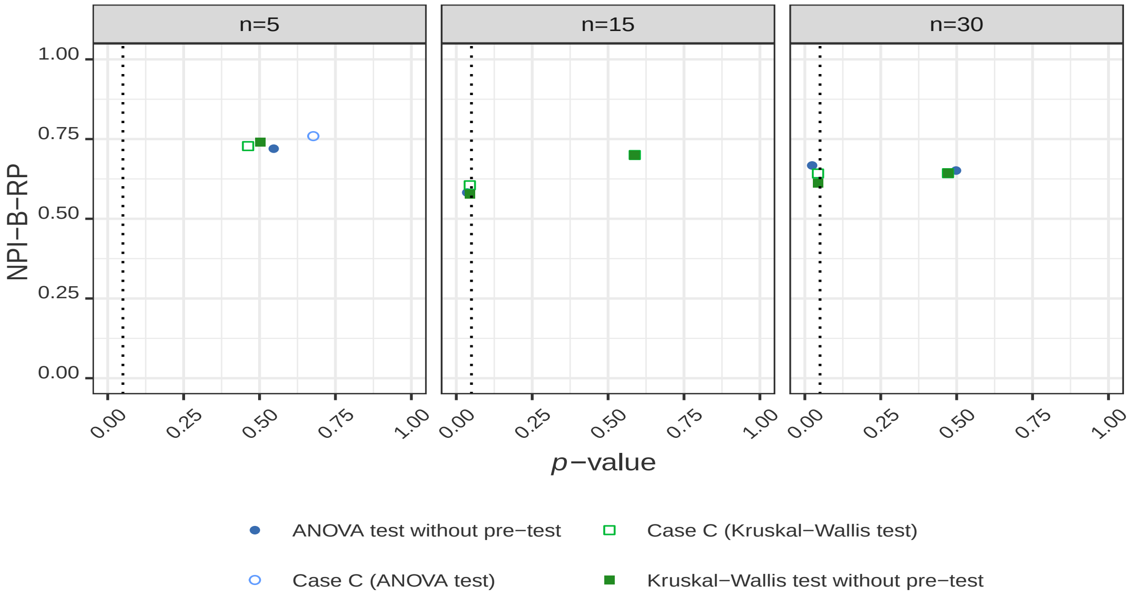


Figure E.19: Comparing RP values for location tests with and without preliminary test (*Case C*), plotted against their corresponding mean p-values. When the original samples are drawn from the mixture of Normal distributions $0.4 \cdot N(5, 1^2) + 0.6 \cdot N(15, 2^2)$, $M = 3$.

E.3 Reproducibility for the three-stage procedure

This section evaluates the reproducibility of the three-stage procedure testing. The procedure included location tests for multiple groups and preliminary tests for Normality and equality of variances. Initially, each group's Normality is tested using the Shapiro-Wilk test. If any sample rejects the Normality assumption, the Kruskal-Wallis test is applied. Otherwise, Levene's test for variance equality is conducted. If the test fails to reject the null hypothesis, one-way ANOVA is performed. However, if the null hypothesis is rejected, indicating unequal variances, Welch's ANOVA is used instead.

Welch's one-way ANOVA is employed when the assumption of equal variances is not met to compare the means of more than two independent groups. It has null and alternative hypotheses like the conventional ANOVA test, which assumes equal population variances across all groups [48].

The steps outlined in Section 5.3 are utilized to assess RP for the three-stage procedure, accounting for changes in tests such as switching from Welch's t-test to Welch's ANOVA and from the F-test to Levene's test. The simulation studies in Section 6.3 are applied for three and five groups, samples are sampled from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$ for three groups, and $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$ for five groups.

The simulation results for the reproducibility of the three-stage procedure testing for *Cases A, B*, and *C*, as well as the reproducibility of multiple-sample location tests without the preliminary test. These findings exhibit a pattern of reproducibility similar to that observed in the two-stage procedure in Section 6.3.1.

Figures E.20 and E.21 show RP for multiple-sample location tests without preliminary tests for 3 and 5 groups, respectively, under the scenario where group distributions are Normal and have the same mean but have different variances. RP values for all location tests show the general pattern of RP. There is variability in RP values and it decreases with increasing in the sample size. Generally, RP values decrease in the non-rejection area and increase in the rejection area with larger sample sizes and numbers of groups. In cases of small sample sizes, RP for the Kruskal-Wallis test is slightly higher than that for the parametric test in the non-rejection area. However, for larger sample sizes, RP for the

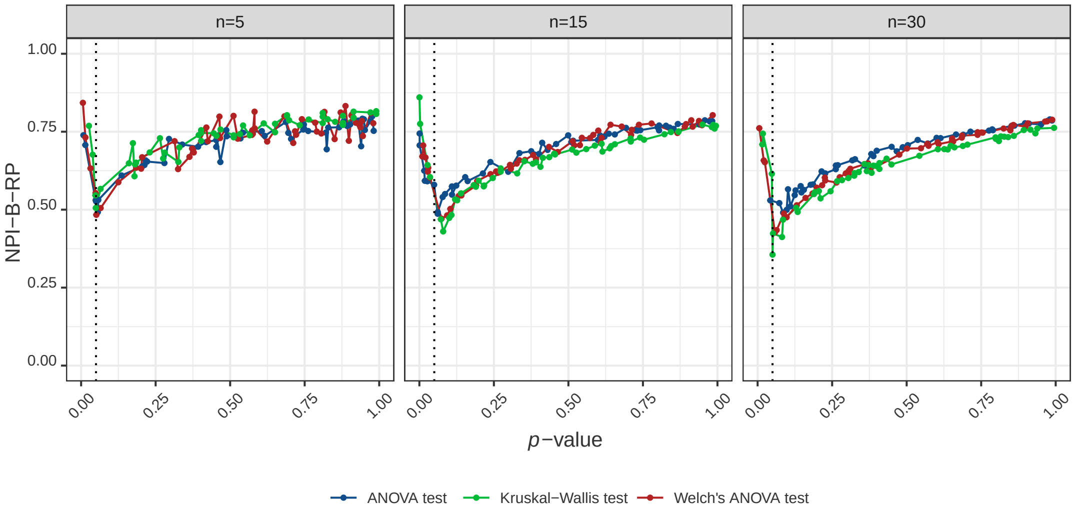


Figure E.20: RP values against the p-values for location tests without preliminary tests, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M=3$.

ANOVA test betters that for Welch's ANOVA and the Kruskal-Wallis test, particularly evident with a higher number of groups. This is because as the sample size increases, the power of the ANOVA test increases, as it becomes more robust to the violation of equal variances assumption. In contrast, the power of Welch's ANOVA and the Kruskal-Wallis test does not increase as dramatically with larger sample sizes [64].

Figures E.22 -E.27 show the simulation studies results for reproducibility in the three-stage procedure for *Case A, B* and *C*. The RP values for both parametric and nonparametric tests follow similar patterns observed in the two-stage procedure for *Case A, B* and *C* presented in Section 6.3.1.

The effect of performing the preliminary tests for the Normality and the equality of variances on RP for the multiple-sample location tests seem small for all cases as shown in Figures E.28 - E.33.

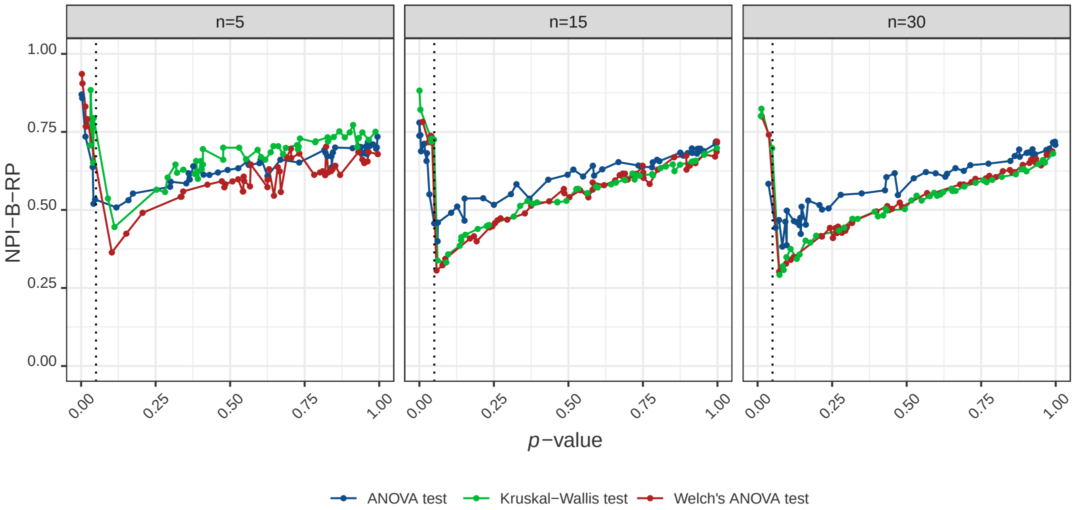


Figure E.21: RP values against the p-values for location tests without preliminary tests, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M = 5$.

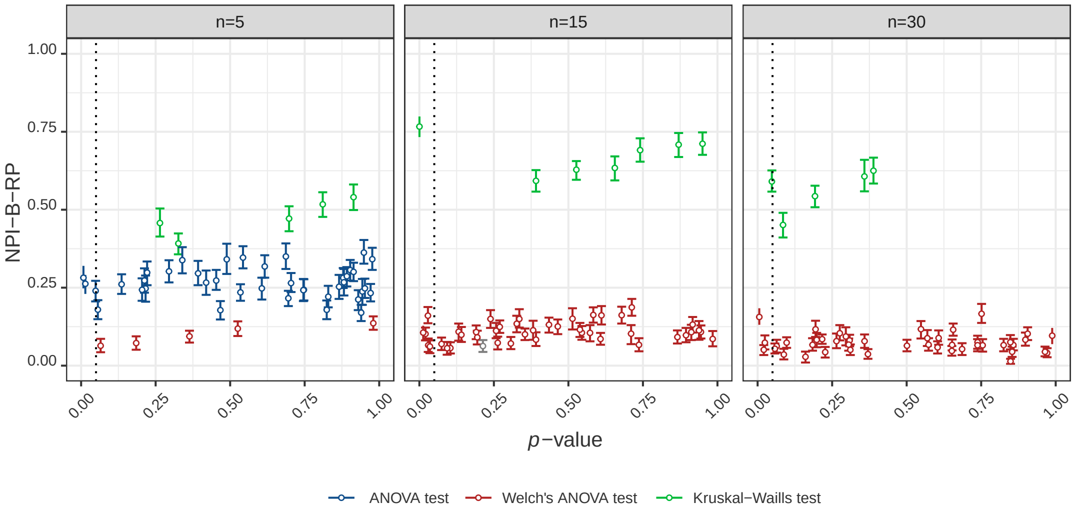


Figure E.22: RP values for three-stage procedure *Case A*, plotted against the p-values for location tests, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M = 3$.

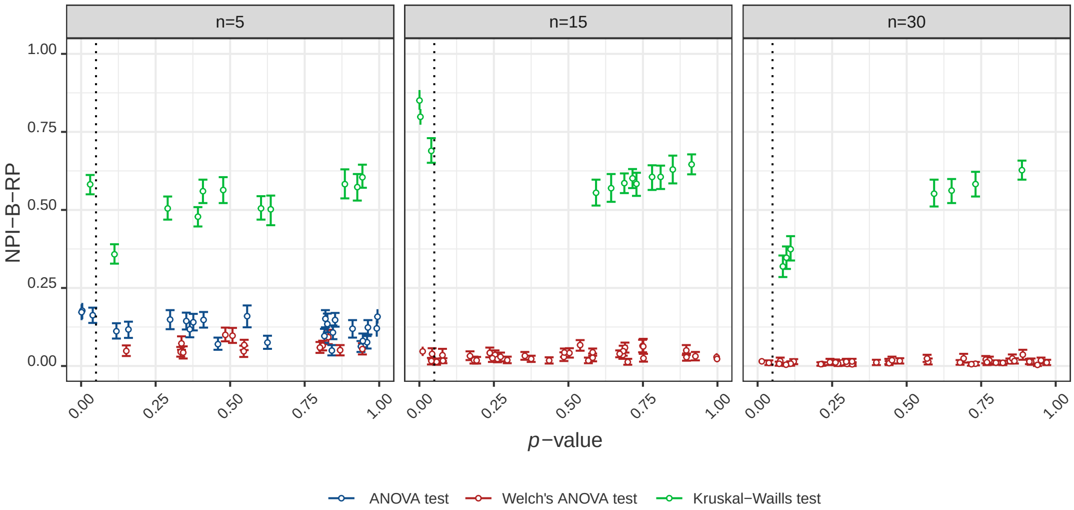


Figure E.23: RP values for three-stage procedure *Case A*, plotted against the p-values for location tests, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M = 5$.

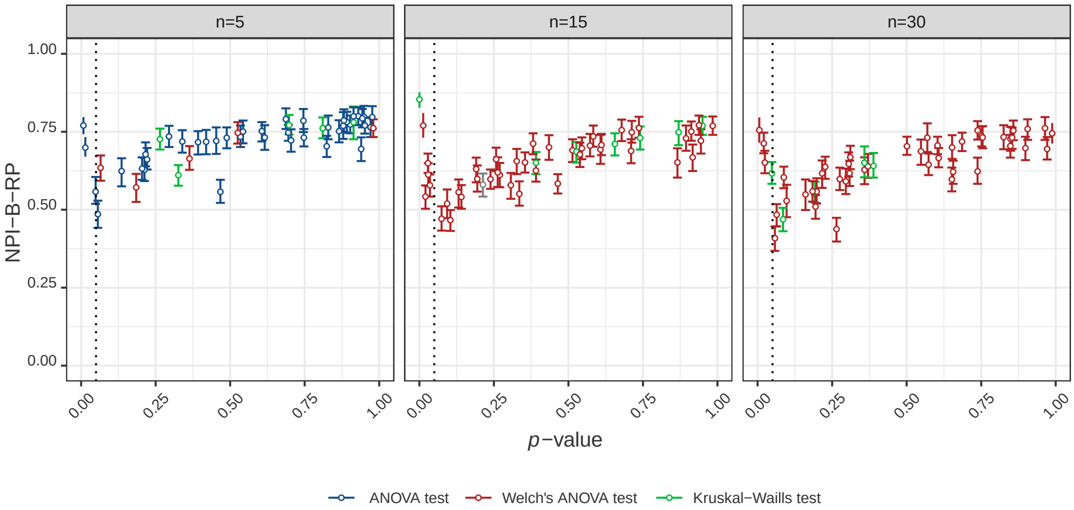


Figure E.24: RP values for three-stage procedure *Case B*, plotted against the p-values for location tests, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M = 3$.

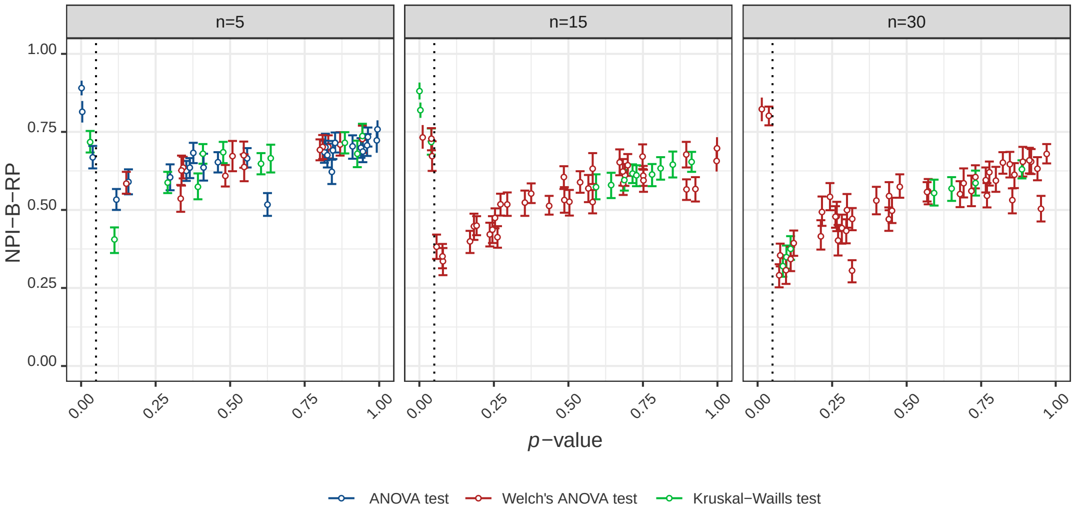


Figure E.25: RP values for three-stage procedure *Case B*, plotted against the p-values for location tests, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M = 5$.

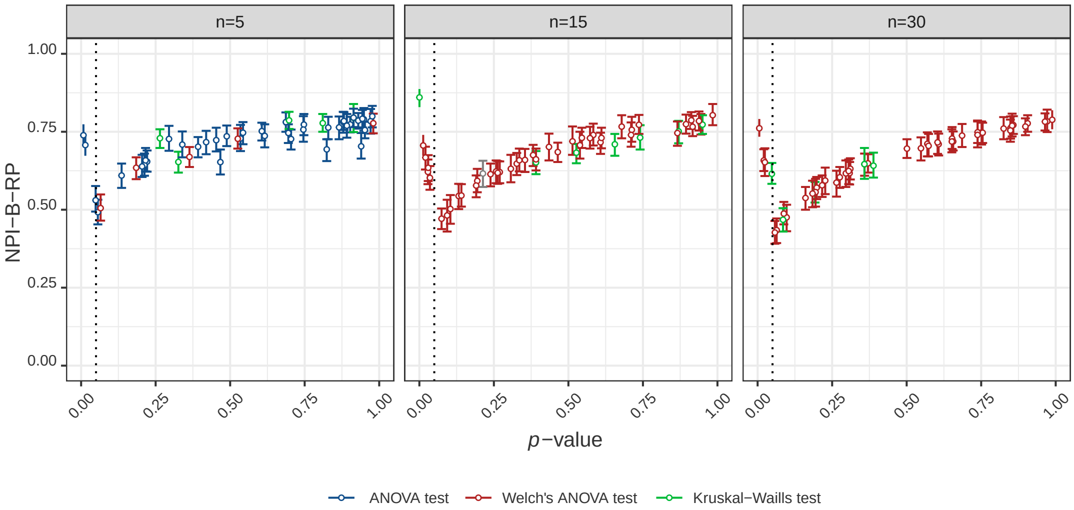


Figure E.26: RP values for three-stage procedure *Case C*, plotted against the p-values for location tests, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M = 3$.

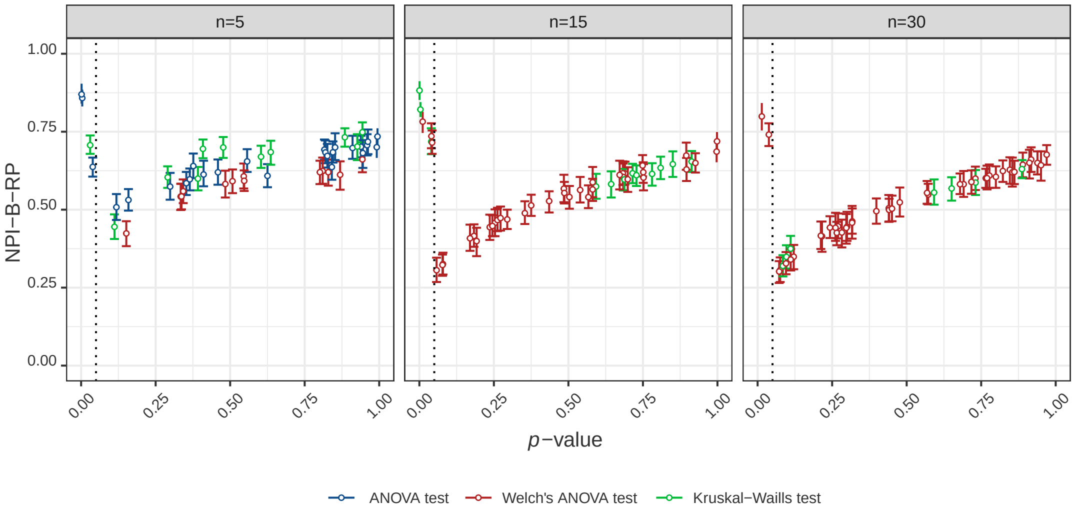


Figure E.27: RP values for three-stage procedure *Case C*, plotted against the *p*-values for location tests, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M = 5$.

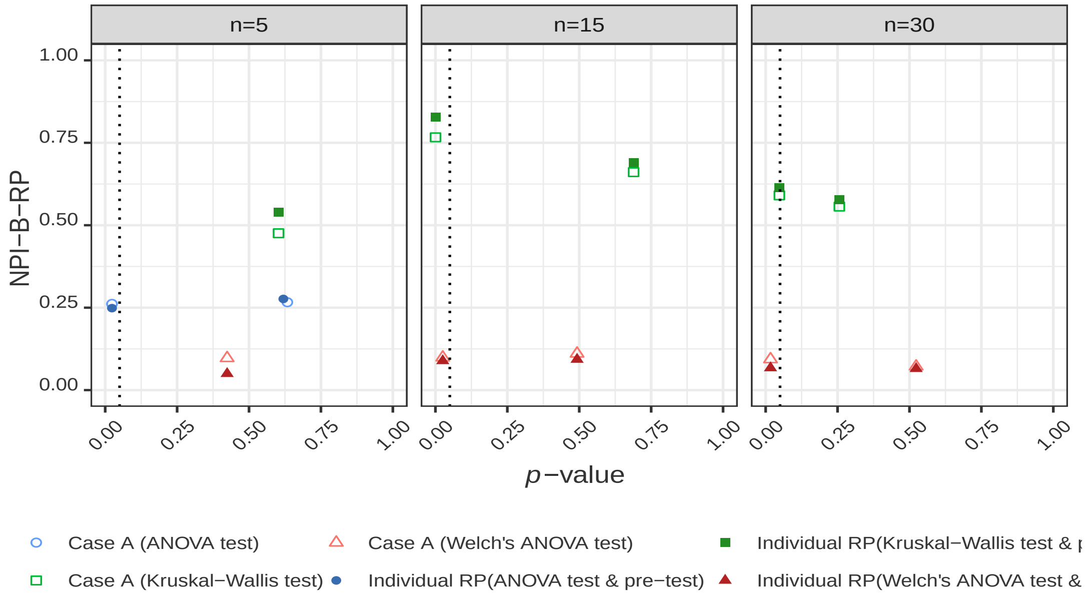


Figure E.28: Comparison of the mean of RP for location tests with and without preliminary tests *Case A*, against the mean of their *p*-values, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M = 3$.

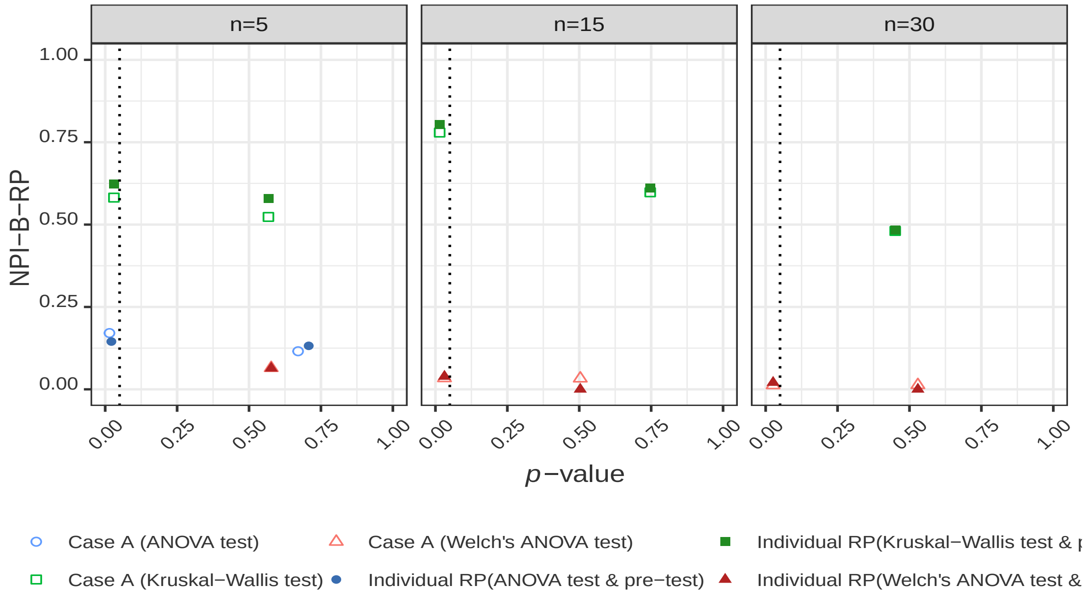


Figure E.29: Comparison of the mean of RP for location tests with and without preliminary tests *Case A*, against the mean of their p-values, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M = 5$.

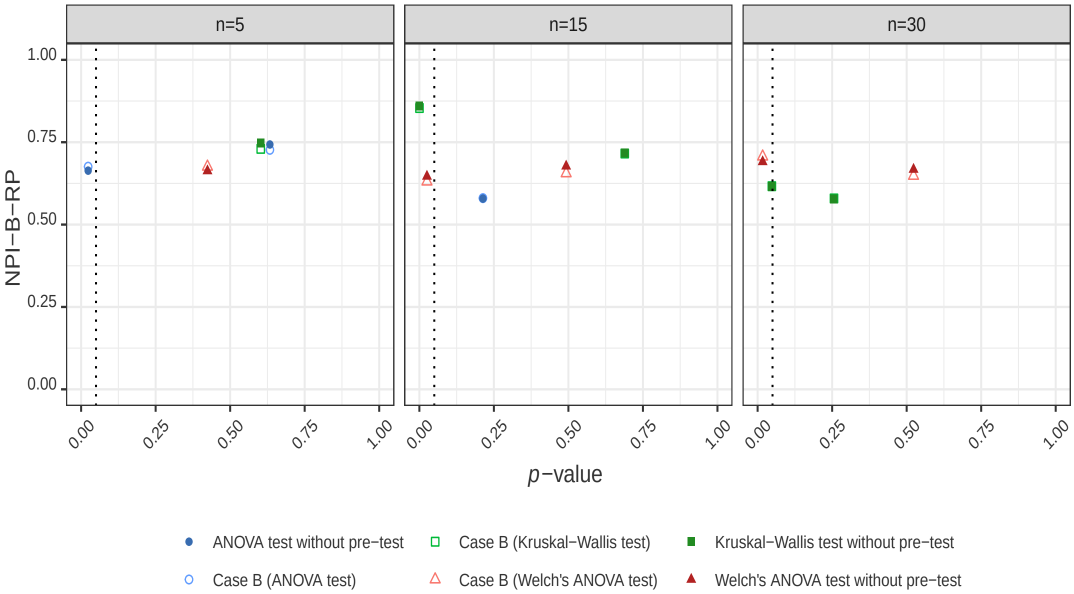


Figure E.30: Comparison of the mean of RP for location tests with and without preliminary tests *Case B*, against the mean of their p-values, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M = 3$.

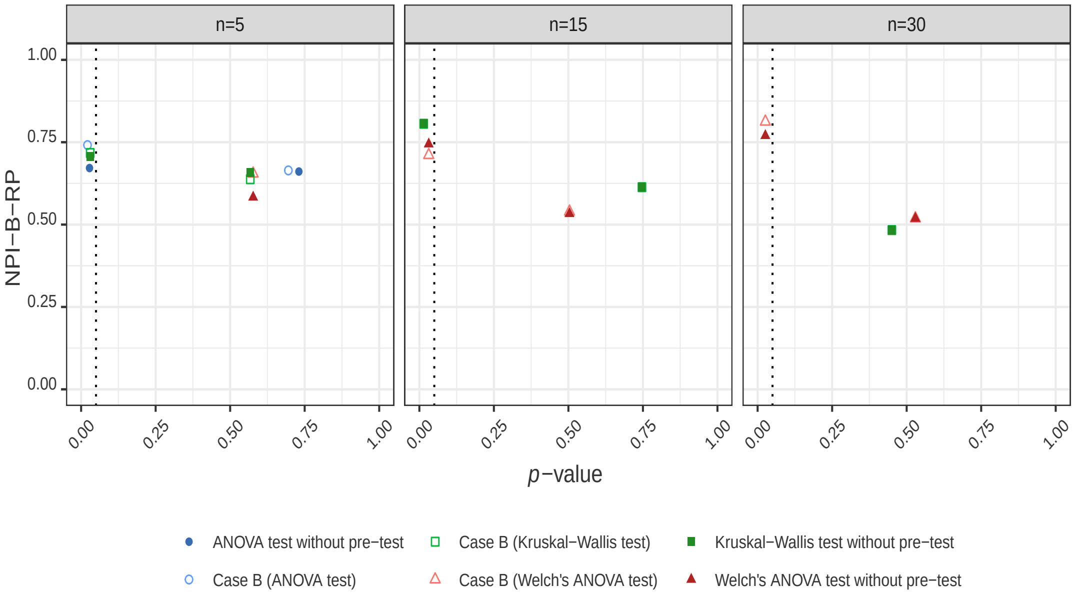


Figure E.31: Comparison of the mean of RP for location tests with and without preliminary tests *Case B*, against the mean of their p-values, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M=5$.

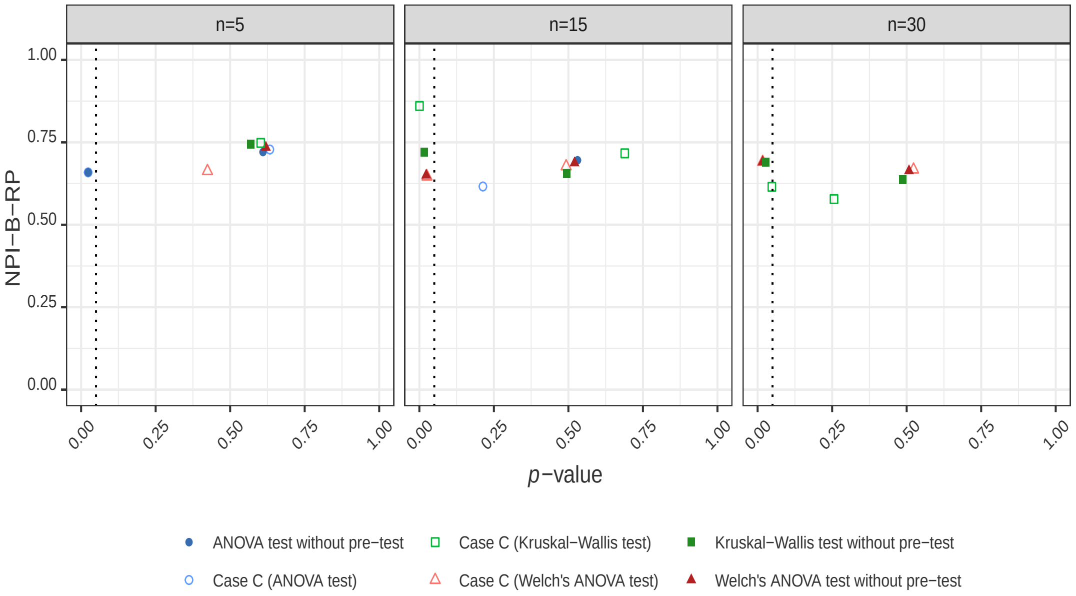


Figure E.32: Comparison of the mean of RP for location tests with and without preliminary tests *Case C*, against the mean of their p-values, samples from $N(0,1)$, $N(0,2^2)$, and $N(0,4^2)$, $M=3$.

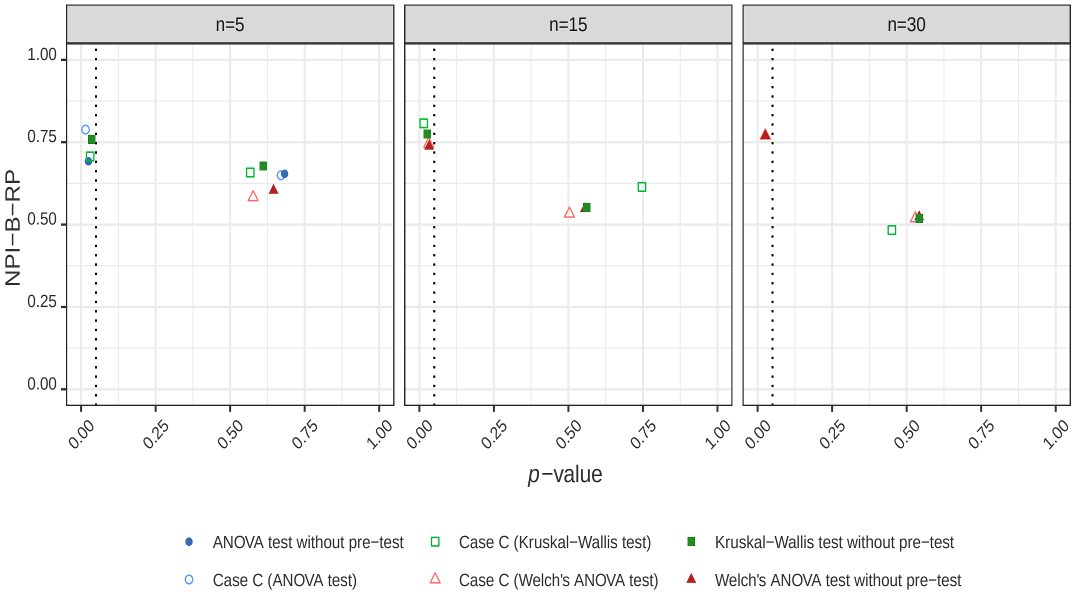


Figure E.33: Comparison of the mean of RP for location tests with and without preliminary tests *Case C*, against the mean of their p-values, samples from $N(0,1)$, $N(0,2^2)$, $N(0,4^2)$, $N(0,1)$, and $N(0,1)$, $M = 5$.

Bibliography

[1] Abdi, H. and Williams, L. J. (2010). Tukey's honestly significant difference (hsd) test. *Encyclopedia of Research Design*, 3(1):1–5.

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

[3] Alqifari, H. N. (2017). *Nonparametric Predictive Inference for Future Order Statistics.* PhD thesis, Durham University.

[4] Altman, D. G. (1990). *Practical statistics for medical research.* CRC Press.

[5] Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics*, pages 193–212.

[6] Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49(268):765–769.

[7] Atmanspacher, H. and Maasen, S. (2016). *Reproducibility: principles, problems, practices, and prospects.* John Wiley & Sons.

[8] Augustin, T., Coolen, F. P. A., De Cooman, G., and Troffaes, M. C. (2014). *Introduction to Imprecise Probabilities.* John Wiley & Sons.

[9] Bares, R., Evans, D., and Long, S. (2010). Noise estimation in long-range matched-filter envelope sonar data. *IEEE Journal of Oceanic Engineering*, 35(2):230–235.

[10] Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, 160(901):268–282.

[11] Baxevanis, A., Bader, G., and Wishart, D. (2020). *Bioinformatics.* Wiley.

[12] Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, 73(sup1):291–295.

[13] BinHimd, S. (2014). *Nonparametric Predictive Methods for Bootstrap and Test Reproducibility.* PhD thesis, Durham University.

[14] Blanca, M. J., Arnau, J., López-Montiel, D., Bono, R., and Bendayan, R. (2013). Skewness and kurtosis in real data samples. *Methodology.*

[15] Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni Del R Istituto Superiore Di Scienze Economiche E Commericiali Di Firenze*, 8:3–62.

[16] Boon, P. C. (1999). *Asymptotic Behavior of Pretest Procedures.* PhD thesis, Universiteit Twente, Enschede, The Netherlands.

[17] Boslaugh, S. (2012). *Statistics in a Nutshell: A Desktop Quick Reference.* O'Reilly Media, Inc.

[18] Box, G. E. (1953). Non-normality and tests on variances. *Biometrika*, 40(3/4):318–335.

[19] Chowdhury, G. G. and Chowdhury, S. (2011). *Information Users and Usability in the Digital Age.* Facet Publishing.

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

[21] Coolen, F. P. A. (2011). *Nonparametric Predictive Inference*, pages 968–970. Springer Berlin Heidelberg, Berlin, Heidelberg.

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

[23] Coolen, F. P. A. and Alqifari, H. N. (2019). Robustness of nonparametric predictive inference for future order statistics. *The Journal of Statistical Theory and Practice*, 13:12.

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

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

[26] Coolen, F. P. A. and Coolen-Maturi, T. (2024). Statistical reproducibility. *International Encyclopedia of Statistical Science, M. Lovric (Ed.). Springer, Heidelberg.*

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

[28] Coolen, F. P. A. and Yan, K. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126(1):25–54.

[29] Coolen, F. P. A. and Yan, K.-J. (2003). Nonparametric predictive comparison of two groups of lifetime data. In *ISIPTA*, volume 3, pages 148–161.

[30] Coolen-Maturi, T., Coolen, F. P. A., and Alqifari, H. (2018). Non-parametric predictive inference for future order statistics. *Communications in Statistics - Theory and Methods*, 47(10):2527–2548.

[31] Cramér, H. (1928). On the composition of elementary errors. *Scandinavian Actuarial Journal*, 1928(1):13–74.

[32] D'agostino, R. B. (1970). Transformation to normality of the null distribution of g1. *Biometrika*, 57:679–681.

[33] Daniel, W. W. and Cross, C. L. (2019). *Biostatistics: A Foundation for Analysis in the Health Sciences.* Wiley.

[34] Davis, G. and Pecar, B. (2013). *Business Statistics Using Excel.* Business Statistics Using Excel. OUP Oxford.

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

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

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

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

[39] De Muth, J. E. (2014). *Basic Statistics and Pharmaceutical Statistical Applications.* CRC Press.

[40] Easterling, R. G. and Anderson, H. E. (1978). The effect of preliminary normality goodness of fit tests on subsequent inference. *Journal of Statistical Computation and Simulation*, 8:1–11.

[41] Freidlin, B., Miao, W., and Gastwirth, J. L. (2003). On the use of the shapiro-wilk test in two-stage adaptive inference for paired data from moderate to very heavy tailed distributions. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 45(7):887–900.

[42] Gibbons, J. and Chakraborti, S. (2010). *Nonparametric Statistical Inference.* Chapman and Hall/CRC, 5th edition.

[43] Glover, T. and Mitchell, K. (2015). *An Introduction to Biostatistics: Third Edition.* Waveland Pres.

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

[45] Grootveld, M. (2014). *Metabolic Profiling: Disease and Xenobiotics.* Number 21. Royal Society of Chemistry.

[46] Gross, J. and Ligges, U. (2015). nortest: Tests for normality. *R package version*, 1(4).

[47] Gurland, J. and McCullough, R. S. (1962). Testing equality of means after a preliminary test of equality of variances. *Biometrika*, 49(3-4):403–417.

[48] Hassouna, A. (2023). *Statistics for Clinicians: How Much Should a Doctor Know?* Springer Nature.

[49] Hawkins, D. (2019). *Biomeasurement.* Oxford University Press.

[50] Hayes, A. W. and Kobets, T. (2023). *Hayes' Principles and Methods of Toxicology.* Crc Press.

[51] Henderson, A. R. (2006). Testing experimental data for univariate normality. *Clinica Chimica Acta*, 366(1-2):112–129.

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

[53] Hoag, J. R. and Kuo, C.-L. (2017). Normal and non-normal data simulations for the evaluation of two-sample location tests. *Monte-Carlo Simulation-Based Statistical Modeling*, pages 41–57.

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

[55] Kennedy, J. J. and Bush, A. J. (1985). *An Introduction to the Design and Analysis of Experiments in Behavioral Research.* University Press of America.

[56] Keselman, H., Othman, A. R., and Wilcox, R. R. (2013). Preliminary testing for normality: Is this a good practice? *Journal of Modern Applied Statistical Methods*, 12(2):2.

[57] Khan, A. and Rayner, G. D. (2003). Robustness to non-normality of common tests for the many-sample location problem. *Advances in Decision Sciences*, 7(4):187–206.

[58] Kolmogorov, A. (1933). Sulla determinizione empirica di una legge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91.

[59] Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004). *Applied Linear Statistical Models*. McGraw-Hill.

[60] Lantz, B., Andersson, R., and Manfredsson, P. (2016). Preliminary tests of normality when comparing three independent samples. *Journal of Modern Applied Statistical Methods*, 15:135–148.

[61] Levene, H. (1960). Contributions to probability and statistics: Essays in honor of harold hotelling. *Palo Alto*, pages 278–292.

[62] Lilliefors, H. W. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, 62(318):399–402.

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

[64] Liu, H. (2015). *Comparing Welch ANOVA, a Kruskal-Wallis test, and traditional ANOVA in case of heterogeneity of variance*. Virginia Commonwealth University.

[65] Marko, S. and Erik, M. (2011). A concise guide to market research the process, data, and methods using ibm spss statistics.

[66] McComb, B., Zuckerberg, B., Vesely, D., and Jordan, C. (2010). *Monitoring Animal Populations and Their Habitats: A Practitioner's Guide*. CRC Press.

[67] Meredith, W. M., Frederiksen, C. H., and McLaughlin, D. H. (1974). Statistics and data analysis. *Annual Review of Psychology*, 25(1):453–505.

[68] Miari, M., Anan, M. T., and Zeina, M. B. (2022). *Single Valued Neutrosophic Kruskal-Wallis and Mann Whitney Tests*. Infinite Study.

[69] Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin*, 105(1):156.

[70] Mohd Razali, N. and Yap, B. (2011). Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests. *Journal of Statistical Modeling and Anlytics*, 2(1):21–33.

[71] Moubray, J. (2001). *Reliability-centered Maintenance*. G - Reference,Information and Interdisciplinary Subjects Series. Industrial Press.

[72] MyoungJin, K., Caroline, M., and Teresa, D. V. (2020). *Statistics for Evidence-Based Practice in Nursing*. Jones & Bartlett Learning.

[73] Navarro, D. (2015). Learning statistics with r: Daniel joseph navarro. *Morrisville, NC: Lulu. com*.

[74] Nelson, M. (2020). *Statistics in Nutrition and Dietetics*. Wiley.

[75] Ntiwa Foudjo, A. (2013). *Robust normality test and robust power transformation with application to state change detection in non normal processes*. PhD thesis, Technische Universität.

[76] Randles, R. H., Fligner, M. A., Policello, G. E., and Wolfe, D. A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, 75(369):168–172.

[77] Rasch, D. and Guiard, V. (2004). The robustness of parametric statistical methods. *Psychology Science*, 46:175–208.

[78] Rasch, D., Kubinger, K. D., and Moder, K. (2011). The two-sample t-test: Pre-testing its assumptions does not pay off. *Statistical Papers*, 52(1):219–231.

[79] Rice, J. A. (2006). *Mathematical Statistics and Data Analysis*. Cengage Learning.

[80] Rochon, J., Gondan, M., and Kieser, M. (2012). To test or not to test: Preliminary assessment of normality when comparing two independent samples. *BMC Medical Research Methodology*, 12(1):1–11.

[81] Royston, P. (1982). An extension of shapiro and wilk's w test for normality to large samples. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 31(2):115–124.

[82] Royston, P. (1992). Approximating the shapiro-wilk w-test for non-normality. *Statistics and Computing*, 2(3):117–119.

[83] Ruscio, J. and Roche, B. (2012). Variance heterogeneity in published psychological research. *Methodology*.

[84] Schucany, W. R. and Tony Ng, H. (2006). Preliminary goodness-of-fit tests for normality do not validate the one-sample student t. *Communications in Statistics-Theory and Methods*, 35(12):2275–2286.

[85] Senn, S. (2002). A comment on a comment on replication p-values and evidence. *Statistics in Medicine*, 21(16):2437–2444.

[86] Shamsudheen, I. and Hennig, C. (2023). Should we test the model assumptions before running a model-based test? *Journal of Data Science, Statistics, and Visualisation*, 3(3).

[87] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

[88] Sheskin, D. J. (2003). *Handbook of Parametric and Nonparametric Statistical Procedures*. Chapman and Hall/CRC.

[89] Siegel, S. and Castellan Jr, N. J. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

[90] Simkus, A. (2023). *Contributions to Statistical Reproducibility and Small-Sample Bootstrap*. PhD thesis, Durham University.

[91] Simkus, A., Coolen, F. P. A., Coolen-Maturi, T., Karp, N. A., and Bendtsen, C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research*, 31(4):673–688.

[92] Sirkin, R. M. (2006). *Statistics for the Social Sciences.* Sage.

[93] Snedecor, G. W. and Cochran, W. G. (1989). Statistical methods, eight edition. *Iowa State University Press, Ames, Iowa*, 1191(2).

[94] Stephens, M. A. (2017). Tests based on edf statistics. In *Goodness-of-fit-techniques*, pages 97–194. Routledge.

[95] Student (1908). The probable error of a mean. *Biometrika*, 6(1):1–25.

[96] Tavakoli, H. (2012). *A Dictionary of Research Methodology and Statistics in Applied Linguistics.* Rahnama.

[97] Tu, W. (2007). Basic principles of statistical inference. *Topics in Biostatistics*, pages 53–72.

[98] van Straelen, R., Kesenne, S., Kesenne, S., and Reyns, C. (2004). *Kwantitatief Bekeken: Liber Amicorum Prof. dr. Robert Van Straelen.* Garant.

[99] Vaughan, L. (2001). *Statistical Methods for the Information Professional: A Practical, Painless Approach to Understanding, Using, and Interpreting Statistics*, volume 367. Information Today, Inc.

[100] Walker, J. and Almond, P. (2010). *Interpreting Statistical Findings: A Guide for Health Professionals and Students: a Guide for Health Professionals and Students.* McGraw-hill education (UK).

[101] Welch, B. L. (1938). The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362.

[102] Wells, C. S. and Hintze, J. M. (2007). Dealing with assumptions underlying statistical tests. *Psychology in the Schools*, 44(5):495–502.

[103] Wyrzykowski, R., Dongarra, J., Deelman, E., and Karczewski, K. (2023). *Parallel Processing and Applied Mathematics: 14th International Conference, PPAM 2022, Gdansk, Poland, September 11–14, 2022, Revised Selected Papers, Part II*, volume 13826. Springer Nature.

[104] Zimmerman, D. W. (2011). A simple and effective decision rule for choosing a significance test to protect against non-normality. *British Journal of Mathematical and Statistical Psychology*, 64(3):388–409.

[105] Zimmerman, D. W. (2014). Consequences of choosing samples in hypothesis testing to ensure homogeneity of variance. *British Journal of Mathematical and Statistical Psychology*, 67(1):1–29.