Nonparametric Predictive Inference For Reproducibility of One-Way Layout Tests

NORAH ALALYANI

Nonparametric Predictive Inference For Reproducibility of One-Way Layout Tests

Norah Oudah Alalyani

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
University of Durham
England

September 2024

Dedicated to

My brother Abdullah, may God rest his soul

My beloved children Abdullah and Noor

My family and friends

Nonparametric Predictive Inference For Reproducibility of One-Way Layout Tests

Norah Oudah Alalyani

Submitted for the degree of Doctor of Philosophy
September 2024

Abstract

The reproducibility of research findings is of main interest in many disciplines. Reproducibility of a statistical test means that, if the experiment were repeated under the same conditions, it would lead to the same conclusion with regard to rejection of the null hypothesis. The probability that the test conclusion for the repeated test would be the same as the original test is called reproducibility probability (RP). The concept of test reproducibility is inherently a predictive inference problem. This thesis investigates the reproducibility of statistical hypothesis tests for One-Way Layout tests using Nonparametric Predictive Inference (NPI). NPI is a predictive approach based on few modelling assumptions that considers multiple future observations that are exchangeable with the data observations which makes it suitable for inference about reproducibility. The uncertainty can be quantified in NPI reproducibility through lower and upper reproducibility probabilities.

This thesis considers reproducibility of general alternatives tests, including the Kruskal Wallis test and the one-way ANOVA test, as well as the Jonckheere-Terpstra test for the ordered alternative hypothesis. This thesis also considers reproducibility probabilities for the umbrella alternatives tests, specifically the Mack-Wolfe test and the Esra-Fikri test, as well as for slippage tests, namely, the Mosteller test. Deriving the exact NPI lower and upper reproducibility probabilities is not trivial for some tests and computationally challenging for large sample sizes. To address these difficulties, two NPI-based approaches are implemented, namely, the NPI sampling of orderings and the NPI-bootstrap techniques. The NPI reproducibility is low when the test statistic is close to the threshold between rejecting and not rejecting the null hypothesis. If the test statistic is close to the rejection threshold for tests with directional alternatives, reproducibility tends to be lower for rejection of the null hypothesis than for non-rejection. This may be problematic, in particular as rejection of the null hypothesis is often the main goal of statistical experiments.

Declaration

The work in this thesis is based on research carried out in the Department of Mathematical Sciences at Durham University. No part of this thesis has been submitted elsewhere for any degree or qualification.

Acknowledgements

I am truly grateful to Allah for the countless blessings he has bestowed on me generally in my life and particularly in accomplishing this thesis.

I would like to express my sincere appreciation and special thanks to my supervisors Prof. Frank Coolen and Dr. Tahani Coolen-Maturi for unlimited support, expert advice and exceptional guidance in all stages of this thesis.

I am very grateful to my friends and colleagues for their support and companionship during my time in Durham, particular thanks go to Kholood Alyazidi, Fatimah Alshihry and Nouf Alawaji for all their support and help. My heartfelt gratitude goes to Amjad Almazroei for her encouragement, friendship, and the memories we have shared throughout this journey. I also wish to thank Amira Alrewetae for her support and kindness.

I am also deeply grateful to my parents and siblings for their endless love and for being there during the hardest times. They have all been a source of motivation. A special note of appreciation and thanks go to my wonderful children Abdullah and Noor for their patience and enthusiasm during my PhD journey.

I would like also to thank my examiners Prof. Malcolm Farrow and Dr. Reza Drikvandi for their valuable discussions.

I gratefully acknowledge the financial support from King Faisal University, Saudi Arabia, and the Saudi Arabian Cultural Bureau in London, enabling me to pursue my PhD studies at Durham University. Many thanks also to Durham University for providing me with an enjoyable academic atmosphere.

My final thanks go to everyone who has assisted me, stood by me or contributed to my educational progress in any way. This thesis is dedicated to all of you.

Notations

NPI	Nonparametric Predictive Inference
$A_{(n)}$	Hill's assumption
RP	Reproducibility Probability
NPI-RP	NPI Reproducibility Probability
NPI-RP-E	NPI reproducibility probability using the exact approach
$\underline{RP}$	NPI lower reproducibility probability
$\overline{RP}$	NPI upper reproducibility probability
NPI-B	NPI Bootstrap
B	The number of bootstrap replications
NPI-RP-B	NPI reproducibility probability using NPI Bootstrap
NPI-RP-SO	NPI reproducibility probability using Sampling of Orderings approach
r^*	The number of orderings sampled in the NPI-RP-SO
$\widehat{\underline{RP}}$	NPI lower reproducibility probability using sampling of orderings
$\widehat{\overline{RP}}$	NPI upper reproducibility probability using sampling of orderings
k	The number of independent groups
n	The sample size for a particular group
N	The total number of observations in the k groups
KW	The Kruskal-Wallis test statistic
F	The one-way Analysis of Variance (ANOVA) test statistic
η^2	The effect size for the one-way Analysis of Variance test
ε^2	The effect size for the Kruskal-Wallis test
J	The Jonckheere-Terpstra test statistic
A_p	The Mack-Wolfe test statistic
$\tilde{A}_p$	The Esra-Fikri test statistic

Contents

Chapter 1

Introduction

1.1 Motivation

Hypothesis testing is a commonly used methodology for comparing two or more groups in statistical inference. Hypothesis tests are generally categorized into parametric and nonparametric tests. Parametric tests such as the t-test and the Analysis of Variance (ANOVA) test require some assumptions about the underlying population distribution and homogeneity of variances. When the assumptions are violated, there are equivalent rank-based nonparametric tests which can be used, such as the Wilcoxon-Mann-Whitney test and the Kruskal-Wallis test. Other nonparametric tests for comparing more than two groups that consider different alternative hypotheses are the Jonckheere-Terpstra test, the Mack-Wolfe test and the Mosteller test.

There has been an increase in interest in reproducibility of results in many scientific fields. A survey conducted by *nature* indicates that there is a reproducibility crises as high percentages of the researchers in the survey have failed to reproduce another scientist's experiments results, and to reproduce their own experiments [8]. The reproducibility of statistical test conclusions is of main interests in statistics. If the test is repeated under identical circumstances would the same conclusion as the original test be reached with regard to rejection or non-rejection of the null hypothesis? The probability that the test conclusion for the repeated test would be the same as the original test is called reproducibility probability (RP) [18]. The reproducibility probability of statistical hypothesis tests is crucial for the reliability of tests outcomes. Goodman [51] started the discussion about the concept of reproducibility of a statistical test conclusion, and indicated that the failure of an experiment to repeat the statistical significance achieved by previous studies often causes concern in the medical literature due to a misunderstanding of the p-value. While this thesis focuses on the reproducibility of statistical hypothesis tests, reproducibility in a wider context is an important topic including the detailed recording of

practical software reproducibility using the same software tools, code, data, and environment as originally used, etc. There is a large number of publications consider the reproducibility issue in various scientific fields, in psychology [68, 87, 100, 110], chemistry [14, 49], computer sciences and machine learning [54, 89]. Such issue have been discussed in detail in the recent thesis by Simkus [97].

In this thesis, the aim is to study the reproducibility probability of statistical hypothesis tests, using Nonparametric Predictive Inference (NPI). NPI is a statistical framework based on the assumption $A_{(n)}$ proposed by Hill for prediction of a future observation [57, 58]. In the NPI approach, uncertainty is quantified via lower and upper probabilities for events of interest. NPI has been introduced for many applications in statistics, reliability, finance, as it is easy to implement and relies on few assumptions [26, 30, 31, 34, 35, 44]. The existing researches have shown that NPI has good statistical properties and gives reliable predictive results.

The reproducibility probability is naturally considered as a predictive problem which aligns with NPI approach. Hence, the NPI approach is suitable for studying reproducibility of a test. First application of NPI to reproducibility probability for statistical hypothesis tests was introduced by Coolen and BinHimd [32], who investigated NPI reproducibility for some basic nonparametric tests: the one-sample sign test, the one-sample Wilcoxon signed rank test, the two-sample rank sum test also known as Wilcoxon-Mann-Whitney test and the two-sample Kolmogorov-Smirnov test. Coolen and Alqifari [29] developed NPI-RP for two classical statistical tests based on order statistics, namely a one-sample quantile test and a precedence test. Marques et al. [75] studied NPI-RP for the likelihood ratio test.

This thesis contributes to the development of NPI reproducibility probability of statistical tests by considering some one-way layout tests, namely, the Analysis of Variance (ANOVA) test, the Kruskal-Wallis (KW) test, the Jonckheere-Terpstra (JT) test, the Mack-Wolfe (MW) test, the Esra-Fikri (EF) test and the Mosteller test.

The outline of this introductory chapter is as follows. Section 1.2 introduces some background information about the one-way layout tests. Section 1.3 presents a brief introduction to the topic of the reproducibility in the literature. In Section 1.4, an overview of the main concept of Nonparametric Predictive Inference (NPI) is introduced. The application of the NPI approach in developing methods to estimate the reproducibility probability for a statistical test, is discussed in Section 1.5. Finally, a detailed outline of this thesis is given in Section 1.6.

1.2 One-way layout tests

There are several statistical tests in the literature to compare multiple groups, e.g. to test if there is a significant difference in the location parameters such as the means or the medians of the groups, these tests are called One-way layout tests. There are different alternative hypotheses. One of the commonly used statistical tests for general alternatives is the Analysis of Variance (ANOVA) test which is used to compare the location parameters for multiple groups to determine if there are any significant differences between them [67]. The application of the ANOVA test requires certain assumptions that must be satisfied. These assumptions are the independence of observations, normality of probability distributions and equality of variances [66]. When the ANOVA test assumptions are not met the Kruskal-Wallis (KW) test which is the nonparametric equivalent to the ANOVA test can be used [59, 93]. The KW test does not make assumptions about normality and homogeneity of variances. There are other nonparametric tests introduced in the literature for comparing the location parameters for multiple groups. The Jonckheere-Terpstra (JT) test is nonparametric statistical test, where the alternative hypothesis that the location parameters are ordered in a specific way (stating the direction of the order to be increasing or decreasing) [61, 101]. Other statistical tests are the umbrella alternatives tests used for the umbrella alternative hypothesis that the location parameters have a peak at population p [16, 45, 71, 76]. Slippage tests are used to determine whether one or more groups have slipped either to the left or to the right, meaning that an unspecified number of observations in one or more groups is smaller or larger than all the remaining observations in the other groups, respectively [19, 25, 79, 80].

1.3 Reproducibility Probability (RP)

Reproducibility is the ability to reproduce research results when the same methods, data, and analysis as the original study are used. The literature demonstrates that there has been a significant amount of research conducted on statistical reproducibility. Reproducibility and statistical reproducibility has become an increasingly important issue in scientific research, as it is crucial for the scientific community to maintain confidence in research findings and to continue advancing knowledge in various fields, such as biomedical research, psychology, and social sciences. Recently, much attention has been paid to the reproducibility of statistical hypothesis tests [6]. Statistical reproducibility addresses the question: If a statistical test were repeated, under the same circumstances, would it lead to the same conclusion with regard to rejection or non-rejection of the null hypothesis? The probability that the test conclusion for

the repeated test would be the same as the original test is called reproducibility probability (RP) [18, 98].

Initial analysis and discussion on the reproducibility were first presented by Goodman [51], who indicated that the failure of an experiment to repeat the statistical significance achieved by previous studies often causes concern in the medical literature due to a misunderstanding of the p-value. Goodman argues that the probability of replicating a statistically significant result is lower than expected; moreover, the p-value may lead to over optimistic interpretations. In a discussion of Goodman's paper, Senn [94] agreed with Goodman that the nature of reproducibility probability (RP) and p-value are distinct, and emphasized the importance of reproducibility of test conclusions. Nevertheless, he disagreed with Goodman's statement that the p-value overstates the evidence against the null hypothesis and he stated two reasons for this. The first reason depends of the frequentist interpretations of the p-value, if the p-value accepted to be the probability of observing a result as extreme or more extreme than the result observed under the null hypothesis or it is the most stringent possible type I error rate that one could achieve and still reject the null hypothesis. The second reason depends on granting the Bayesian claim that p-values are interpreted by everybody as if they were Bayesian posterior probabilities. Thus, it is not necessarily true that p-value overstates the evidence against the null hypothesis.

Goodman [51] and Senn [94] presented a straightforward argument related to the estimation of RP. They argued that, if the distribution of the test statistic under the null hypothesis is about symmetric, a worst-case scenario could result in an RP of about 0.5. This argument is based on the possibility that the original test statistic value could be equal to the test's critical value. In the absence of additional information, one could anticipate that repeating the experiment would produce a second test statistic value with an equal chance of being larger or smaller than the original value. Consequently, the same conclusion would be reached with probability 0.5. Goodman [51] provides evidence for this claim using a Bayesian approach with a non-informative prior.

Shao and Chow [95] discussed reproducibility probability in the context of clinical trials, where it is a common research concern to assess whether clinical trials that have yielded significant clinical results provide sufficient evidence to guarantee that the findings can be reproduced in a future clinical trial under the same study conditions. Shao and Chow [95] explore the application of three approaches to assess the reproducibility probability for two-sample t-tests: a common power approach where they define the reproducibility probability as the estimated power of the future trial using the data from the previous trial(s); a confidence bounds approach

where RP is defined as a lower confidence bound of the estimated power of the second trial; and a Bayesian approach through the use of the posterior predictive distribution.

De Martini [40] discussed the reproducibility probability estimation of a statistically significant result for parametric tests with one-sided and two-sided alternatives. Specifically, he considered the estimated power of the test and the lower confidence bound of the power as methods for estimating the reproducibility probability, and defined statistical tests based on the reproducibility probability estimation. De Capitani and De Martini [38] considered several estimators for the reproducibility probability for the Wilcoxon Rank Sum test. A comparison between the use of RP and p-value is discussed by De Capitani [37], who comes to the conclusion that the RP defines a decision rule as the p-value. The threshold for defining statistical tests based on the point estimator of the RP turned out to be 0.5, that is, not reject the null hypothesis if the RP estimate is lower than, or equal to 0.5, and reject the null hypothesis otherwise. De Capitani and De Martini [39] provide further discussion on reproducibility probability for the Sign test, the Binomial test, the Kendall test and the Wilcoxon Signed Rank test.

Boos and Stefanski [21] study the variability of p-values in order to gain a deeper understanding of its significance and possible impacts in relation to reproducibility. They use of bootstrap studies, which showed that p-values exhibit surprisingly larger variability than anticipated in typical data situations. Boos and Stefanski [21] extended Shao and Chow's [95] discussion about reproducibility probability using the estimated power approach for t-test to the one-way ANOVA test and found that all $\widehat{RP}$ estimates are relatively close when the p-value equal 0. Boos and Stefanski [21] showed that ANOVA with p-value $= 0.001$ corresponds to $\widehat{RP} = 0.9$, which means that probability of getting p-value < 0.05 and near to 0.001 in a replication of the original experiments is 90%.

Miller [77] emphasized the importance to distinguish between two scenarios for the replication probability; a scenario in which the researchers will obtain significant effects in the repeated experiments where conditions may vary with regard to the original experiments, and the other form of repetition would be obtained by an individual researcher under the same conditions as the original experiments. The inference regarding RP outlined in this thesis aligns entirely with Miller's concept of the 'individual form of repetition'. Killeen [62] considered both of the two scenarios but he did not emphasized the distinction between them.

Killeen [62] emphasized the explicit prediction of reproducibility probability and links it to the effect size. Killeen [62] defined the probability of replication of an experiment results as the probability of finding an effect of the same sign in a replication of an experiment as that found in the original experiment. Although there is some confusion about reproducibility probability

concept in Killeen's paper, the general idea concerning RP exhibits a close alignment with those presented in this thesis, namely a predictive nature of reproducibility probability.

Lecoutre et al. [69] discuss Killeen's [62] approach for probability of replication of experiment result, and consider it to be a "fiducial Bayesian predictive approach" based on noninformative priors. Lecoutre et al. [69] investigated statistical prediction scenarios, for instance, given a significant result in the original experiment estimate, they consider the probability that the result of a replication of the experiment would be significant. Lecoutre et al. [69] found that there remains some confusion in computing the probability of replication, and emphasized the role of predictive procedures in statistical methodology. As stated by Lecoutre et al. [69], predictive probabilities are an unavoidable part of the statistical thinking to help researchers ask and answer experiment-related questions such as "what would happen if additional subjects were to be included into the experiment?", "what would be the conclusion for the data of these future subjects?", or "what would happen if this experiment were to be repeated?".

Billheimer [17] points out that most statistical analyses use hypothesis tests or parameters estimation to form inferential conclusions. However, predictive inference which focuses on predicting future observations given the current data and other related information, can provides multiple advantages for researchers to predict what is likely to happen in future experiments [17].

A comprehensive collection of articles that explore the concept of reproducibility in scientific research are included in the volume edited by Maasen and Atmanspacher [70]. The edited volume covers various topics and principles related to reproducibility, including the definition of reproducibility, reproducibility of experiments or observations that are supported by statistical significance tests, the challenges and problems associated with achieving reproducibility, and the potential benefits of reproducibility for scientific studies. The chapters of this edited volume also discusses different approaches and practices that researchers can adopt to improve the reproducibility of their research. In addition, it includes examples from different scientific fields such as physical sciences, life sciences and social sciences, to demonstrate the importance of reproducibility in advancing scientific research.

The reproducibility is naturally considered as predictive inference problem. Coolen and BinHimd [32] introduced reproducibility probability (RP) as a prediction problem, using Nonparametric Predictive Inference (NPI), and denoted by NPI-RP. This thesis contributes to the statistical reproducibility by considering the NPI-RP for One-way layout tests. An overview of NPI is provided in the following section. NPI methods for test reproducibility which are considered in this thesis will be discussed in details in Section 1.5.

1.4 Nonparametric Predictive Inference (NPI)

Nonparametric Predictive Inference (NPI) is a statistical framework based on $A_{(n)}$ assumption proposed by Hill [57, 58], which provides direct probabilities for future observations given n observations of related random quantities. Inferences based on the assumption $A_{(n)}$ are predictive and nonparametric, and seem suitable if there is hardly any knowledge about the random quantities of interest, other than the n observations, or if one does not want to use such information. Inferences based on such restricted assumptions have also known as 'low structure inferences' [48].

Suppose that $X_1, X_2, \ldots, X_n$ are continuous and exchangeable random quantities. Let the ordered of the observations of $X_1, X_2, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n$, and let $x_0 = -\infty$ and $x_{n+1} = \infty$ for ease of notation. These n observations partition the real-line into $n+1$ intervals $I_j = (x_{j-1}, x_j)$, for $j = 1, ..., n+1$. Given the n observations, the assumption $A_{(n)}$ for the future observation X_{n+1} is:

$$P(X_{n+1} \in (x_{j-1}, x_j)) = \frac{1}{n+1} \qquad \text{for} \quad j = 1, ..., n+1 \tag{1.1}$$

Note that under $A_{(n)}$ it is assumed that ties do not occur. In the NPI framework, tied observations can be dealt with by assuming that such observations differ by a very small amount [58]. In this thesis, the `jitter` function in R is used when relevant.

$A_{(n)}$ does not assume anything else, and can be considered to be a post-data assumption related to exchangeability [47]. $A_{(n)}$ is not sufficient to derive precise probabilities for many events of interest, but optimal bounds based on $A_{(n)}$ can be derived for all events of interest involving X_{n+1}. These bounds are lower and upper probabilities in the theories of imprecise probability [107] and interval probability [109]. Imprecise probability generalizes classical probability in the sense that it can be used for describing uncertainty about events via intervals, instead of a single number. For event A the lower probability is denoted by $\underline{P}(A)$ and the upper probability by $\overline{P}(A)$, with $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, and $\triangle(A) = \overline{P}(A) - \underline{P}(A)$ is called the imprecision [27]. Augustin and Coolen [7] introduced the NPI lower and upper probabilities for the event $X_{n+1} \in B \subset \mathbb{R}$, based on the assumption $A_{(n)}$, as follows:

$$\underline{P}(X_{n+1} \in B) = \frac{1}{n+1} \sum_{j=1}^{n+1} 1\{I_j \subseteq B\} P(X_{n+1} \in I_j) \tag{1.2}$$

$$\overline{P}(X_{n+1} \in B) = \frac{1}{n+1} \sum_{j=1}^{n+1} 1\{I_j \cap B \neq \emptyset\} P(X_{n+1} \in I_j) \tag{1.3}$$

where $1\{A\}$ is an indicator function which is equal to 1 if event A occurs and 0 else. The lower probability $\underline{P}(X_{n+1} \in B)$ can be obtained by counting the probability masses assigned

to interval I_j that are completely within B. The upper probability $\overline{P}(X_{n+1} \in B)$ is the total probability masses that could possibly be within B. The NPI approach has been developed for many applications in statistics, reliability, finance and operational research, as it is easy to implement and relies on few assumptions. NPI has been presented for a range of data types, such as Bernoulli data [26], lifetime data [34, 35], ordinal data [44], bivariate data [81] and multinominal data [30, 31].

The NPI approach can be generalized for $m \geq 1$ future observations, X_{n+i} for $i = 1, ..., m$, based on data consisting of n observations. The data and the future observations are linked via Hill's assumption $A_{(n)}, A_{(n+1)}, ..., A_{(n+m-1)}$, which can be considered as a post-data version of a finite exchangeability assumption for $n+m$ random quantities. [57]. Each future observation is equally likely to fall in any interval I_j between two consecutive observations x_{j-1} and x_j, and all possible orderings of the m future observations among the n data observations are equally likely. There are $\binom{n+m}{n}$ possible orderings O_i where $i = 1, ..., \binom{n+m}{n}$, so the probability of any specific ordering of the m future observations among the n data observations is $\binom{n+m}{n}^{-1}$. Let S_j denote the number of the future observations in the interval $I_j = (x_{j-1}, x_j)$ for $j = 1, ..., n+1$ [26, 28], then inferences about these m future observations can be based on the probabilities

$$P\Big(\bigcap_{j=1}^{n+1} \{S_j = s_j\}\Big) = \binom{n+m}{n}^{-1} \tag{1.4}$$

for any $(S_1, ..., S_{n+1})$ with S_j non-negative integers with $\sum_{j=1}^{n+1} S_j = m$ [28].

In the NPI approach, the lower probability for an event of interest is derived by counting all orderings for which it must hold, whereas the corresponding upper probability is derived by counting all orderings for which it can hold [9]. NPI for multiple future observations has been used for many applications [1, 29].

NPI for m real-valued future observations has been used to study the reproducibility of statistical tests in [4, 18, 74, 98], more details will be given in Section 1.5. In the NPI framework, test reproducibility is viewed as a prediction problem. NPI is a good approach to study statistical test reproducibility since it is based on predicting future observations. In this thesis, NPI for m future observations will be used to derive the NPI lower and upper reproducibility probabilities for the Mack-Wolfe test and the Mosteller test.

1.5 NPI for Reproducibility Probability (NPI-RP)

Recently, the reproducibility of statistical test outcomes has received increasing attention. In particular, the concept of reproducibility probability (RP), where several studies in the literature have suggested different approaches for estimating the reproducibility probability based

on various scenarios. The reproducibility probability (RP) is the probability that the same test result would be reached if the test is repeated under similar conditions [18, 97]. In the NPI approach, the reproducibility probability is considered as a prediction problem. NPI is a frequentist statistics framework that considers m future observations that are exchangeable with given n data observations. Thus, NPI has a predictive nature which makes it suited for studying the reproducibility of a statistical test. In the NPI approach, the attention is restricted to the case where the number of future observations is equal to the number of data observations ($m = n$) which is considered a logical assumption in order to study reproducibility.

NPI for reproducibility probability is introduced by Coolen and BinHimd [32], denoted by NPI-RP. NPI enables inference on test reproducibility by deriving lower and upper probabilities for the event that a repeated test under similar conditions as the original test leads to the same conclusion as the original test, in terms of rejection or non rejection of the null hypothesis [18]. The NPI lower and upper reproducibility probabilities are denoted by $\underline{RP}$ and $\overline{RP}$, respectively. Using the NPI approach, BinHimd [18] investigated the reproducibility probability for some basic nonparametric tests: the one-sample sign test, the one-sample Wilcoxon signed rank test, the two-sample rank sum test also known as Wilcoxon-Mann-Whitney test and the two-sample Kolmogorov-Smirnov test. BinHimd [18] was able to derive the exact NPI lower and upper reproducibility probabilities for some of these tests, then they have been compared to NPI-RP estimates using NPI-Bootstrap (NPI-B). The application of the NPI-B method to study NPI reproducibility probability will denoted with NPI-RP-B, which will be introduced in Section 1.5.3. However, BinHimd [18] found that it is computationally challenging to derive the exact NPI-RP for the Kolmogorov-Smirnov test, and NPI-RP-B is applied to get approximation results for the NPI-RP.

Coolen and Alqifari [29] developed NPI-RP for two classical statistical tests based on order statistics, namely a one-sample quantile test and a precedence test which is used to compare two groups of lifetime data, where one wishes to reach a conclusion before all units on test have failed. This involves considering the NPI approach for future order statistics [4], to derive the NPI lower and upper reproducibility probabilities for the quantile and precedence tests .

Simkus et al. [98] provided an NPI algorithms to study NPI-RP for the t-test as part of pharmaceutical research experiment, using the NPI-RP-B method which provides a point estimate of reproducibility probabilities. Simkus et al. [98] studied the reproducibility probability for an approach that involves multiple pairwise comparisons of groups whose members are given an increasing concentration of a drug. The aim of the experiment is to decide what concentration of the drug is most effective.

Marques et al. [75] studied the NPI reproducibility of the likelihood ratio test using NPI for multiple future observations, which was presented in Section 1.4. Marques et al. [75] proposed a method to derive the exact NPI lower and upper reproducibility of likelihood ratio tests, considering all the possible orderings of the future observations among the data observations. In this methodology, the computation of the exact NPI lower and upper reproducibility probabilities is time consuming when large samples are considered, due to the increase in the number of orderings. To overcome this problem, Marques and Coolen [74] introduced the sampling of orderings technique, where only a random sample of the orderings is considered to estimate the NPI lower and upper reproducibility probabilities.

Alghamdi [3] investigated the reproducibility of statistical hypothesis tests based on randomised response data using the one-sided and the two-sided test, to address the question that for a future test of qualitative randomised response data will the test lead to the same conclusion as the original test?. Alghamdi [3] also considered the reproducibility for estimates in terms of the difference between actual estimates and estimates based on a future data set and developed a measure of reproducibility probability to compare different randomised response methods that can be based on either the NPI lower or upper reproducibility probability.

Aldawsari [2] proposed a new bootstrap method, the parametric predictive bootstrap (PP-B). This method is completely based on parametric models and it is mainly designed for inferences aimed at prediction. Aldawsari [2] applied the PP-B method to study the reproducibility of four parametric tests: one-sample t-test, two sample t-test, Welch's t-test and F-test, as well as comparing its performance with the NPI-RP-B. Aldawsari [2] found that for small sample sizes with the test statistic tending to lie in the rejection region, the RP estimates using PP-B tends to be lower in the case of non-rejection compared to NPI-RP-B, and tends to be higher in the case of rejection than NPI-RP-B. However, increasing the sample size tends to reduce the differences in the RP estimates between PP-B and NPI-RP-B.

In this thesis, the NPI approach for test reproducibility is developed for One-way layout tests. In the following sections, three approaches will be considered: the exact NPI-RP approach, the NPI bootstrap and the sampling of orderings approaches, to compute NPI-based reproducibility probability. The exact NPI-RP analytically provides the NPI lower and upper reproducibility probabilities. Nevertheless, it is computationally challenging to derive the exact NPI lower and upper reproducibility probabilities for large sample sizes and for some statistical tests. The NPI bootstrap and the sampling of orderings methods are both approximation methods for the NPI-RP when it is not possible to compute the NPI-RP exactly. In the NPI-RP-B method, reproducibility probabilities are given as a point estimate, while in the sampling of orderings

method, reproducibility probabilities are given as lower and upper estimates.

1.5.1 Exact NPI-RP (NPI-RP-E)

NPI uses imprecise probability theory that quantify uncertainty by representing probabilities as intervals rather than single point estimates. This section presents the NPI approach for deriving the lower and upper reproducibility probabilities. To demonstrate the method, we utilize the sign test setting, which is the most basic nonparametric test. Let the order of one-sample observations $X_1, \ldots, X_n$, from a distribution with median θ_0, be denoted by $x_1 < x_2 < \ldots < x_n$. For the one sided upper tail test, the hypothesis of interests are $H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$. The test statistic W is the number of observations X_j that are positive,

$$W = \sum_{j=1}^{n} I\{X_j > 0\} \tag{1.5}$$

where $\mathbf{1}\{A\}$ is an indicator function which is equal to 1 if the event A occurs and 0 otherwise. The null hypothesis is rejected at the level of significance α, if $W \geq w_\alpha$, where w_α denotes the upper α percentile for the Binomial distribution with sample size n and probability of success 1/2. NPI approach introduced in Section 1.4 can be applied to make inference about the m future observations among the n data observations. There are $\binom{n+m}{n}$ possible orderings O_i of the m future observations among the n data observations, and all equally likely, where $i = 1, ..., \binom{n+m}{n}$. The reproducibility probability for a statistical test is the probability that for a repeated test the same test outcome would be reached. Hence, in this thesis we restrict attention to the situation with the number of future observations m is equal to the number of data observations n which is considered a logical assumption in order to study reproducibility to maintain consistency and comparability of results, however, the NPI approach can be used for any m. The aim is to find the minimum and maximum for the test statistic W for each ordering O_i, which are denoted by $\underline{W_i}$ and $\overline{W_i}$, respectively.

If the original test conclusion is rejection of H_0, then the NPI lower reproducibility probability is derived by counting the number of orderings for which $\underline{W_i} \geq w_\alpha$. The corresponding NPI upper reproducibility probability is derived by counting the number of orderings for which $\overline{W_i} \geq w_\alpha$. Thus, the NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{\binom{2n}{n}} \sum_i \mathbf{1}\{\underline{W_i} \geq w_\alpha\} \tag{1.6}$$

$$\overline{RP} = \frac{1}{\binom{2n}{n}} \sum_i \mathbf{1}\{\overline{W_i} \geq w_\alpha\} \tag{1.7}$$

where $i = 1, 2, ..., \binom{2n}{n}$. Similarly, if the original test conclusion is non-rejection of H_0 such that $W < w_\alpha$, the NPI lower reproducibility probability is derived by counting the number

of orderings for which $\overline{W_i} < w_\alpha$. The corresponding NPI upper reproducibility probability is derived by counting the number of orderings for which $\underline{W_i} < w_\alpha$. The NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{\binom{2n}{n}} \sum_i \mathbf{1}\{\overline{W_i} < w_\alpha\} \tag{1.8}$$

$$\overline{RP} = \frac{1}{\binom{2n}{n}} \sum_i \mathbf{1}\{\underline{W_i} < w_\alpha\} \tag{1.9}$$

This method to derive the lower and upper reproducibility probabilities for statistical tests is suitable for small sample sizes and requires that deriving $\underline{W_i}$ and $\overline{W_i}$ is relatively easy, hereafter we refer to as the NPI-RP-E. The minimum and the maximum can be derived easily for some test statistics such as the Sign test while it is challenging to derive for other test statistics such as the t-test and the ANOVA test because it is difficult to minimize and maximize the mean and the variance simultaneously, as the variance depends on the mean [18, 97].

1.5.2 NPI-RP using Sampling of Orderings (NPI-RP-SO)

The calculation of the exact NPI lower and upper reproducibility probabilities is computationally expensive for large samples due to the increase in the number of orderings of future observations. One way of implementing NPI-RP for large samples is the sampling of orderings denoted by NPI-RP-SO, which is introduced by Marques and Coolen [74], to get approximations of the NPI lower and upper reproducibility probabilities. In this method, the estimation of $\underline{RP}$ and $\overline{RP}$ following the standard theory of estimation where the sampling procedure of the orderings satisfies the simple random sampling conditions, as there is an equal chance of each ordering to be selected and included in the sample, and the selection of an ordering is independent of the other selections. It should be noted that, when n is not too small, the total number of orderings is large which leads to ignore any possible differences between sampling with and without replacement. The standard theory of estimation of proportions enables us to determine a suitable size for the sample of orderings, depending on a required accuracy of the estimates. The coverage probabilities of the $(1-\alpha)100\%$ confidence intervals for proportions, p, using Normal approximation confidence intervals:

$$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/r^*} \tag{1.10}$$

where $\hat{p}$ is the estimate of the lower or the upper reproducibility probability, r^* is the number of orderings sampled, and $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution. When computing the Normal approximation confidence interval, for some cases where $\hat{p}$ is close to 0 or 1, the lower bound can be less than 0 or the upper bound greater than 1. Thus, the

exact $(1-\alpha)100\%$ confidence interval for $\hat{p}$ is used and the interval bounds are defined by the following [78], the lower bound :

$$P_L = \frac{s}{s + (n-s+1)F_{\alpha/2}(2(n-s+1), 2s)} \tag{1.11}$$

where s is the number of successes in the n Bernoulli trials and $F_{\alpha/2}(2(n-s+1), 2s)$ is the upper $\alpha/2$ th percentile for the F distribution with $2(n-s+1)$ degrees of freedom in the numerator and $2s$ degrees of freedom in the denominator. The upper bound:

$$P_U = \frac{(s+1)F_{\alpha/2}(2(s+1), 2(n-s))}{n-s+(s+1)F_{\alpha/2}(2(s+1), 2(n-s))} \tag{1.12}$$

where $F_{\alpha/2}(2(s+1), 2(n-s))$ is the upper $\alpha/2$ th percentile for the F distribution with $2(s+1)$ degrees of freedom in the numerator and $2(n-s)$ degrees of freedom in the denominator [78].

1.5.3 NPI-RP using NPI-Bootstrap (NPI-RP-B)

Bootstrap methods were introduced in statistics to provide an alternative approach to traditional inference methods, which mainly rely on assumptions about the population distribution. In general, bootstrap methods are relatively easy to implement and can be applied to various statistical problems, including parameter estimation, hypothesis testing and constructing confidence intervals. Some of the known bootstrap methods in statistics are Efron's bootstrap [43], Bank's bootstrap [10] and parametric predictive bootstrap [2].

Efron [43] proposed the standard bootstrap technique, and has since become a widely used for the estimation problems in a variety of applications. Efron bootstrap provides a simple and intuitive way to estimate the sampling distribution of an estimator by resampling from the original data, in situations where the distribution of the estimator is unknown. Banks [10] proposed smoothed version of bootstrap where the empirical cumulative distribution function of the original sample smoothed by linear interpolation, histospline smoothing, between the jump points. Aldawsari [2] proposed a new bootstrap method, the parametric predictive bootstrap, which is completely based on parametric models and it is mainly designed for inferences aimed at prediction.

Nonparametric predictive inference bootstrap (NPI-B) is one of the bootstrap methods which is based on repeated application of Hill's assumption $A_{(n)}$. Coolen and BinHimd [32] developed the NPI-B method for predictive inference, and it does not relies on an assumed parametric model. The performance of the NPI-B method compared with classic bootstrap methods was initially evaluated by Coolen and BinHimd [32, 33], by calculating the variance of statistics, bias, absolute error and mean square error. BinHimd [18] found that the NPI-B method has higher variation than other bootstrap, because the variance of statistics in the NPI-B is closer

to the variance of the original sample than the other bootstrap methods. This is because NPI-B sample observations from both the original sample and the intervals between them. The NPI-B method did not perform well in estimation of confidence intervals [18]. However, it worked well when used for prediction intervals, as it has the best coverage probability [18, 33].

In the NPI-B method, one interval is randomly selected from the $n+1$ intervals created by the n original data observations, and from this interval one future value is drawn uniformly. The first drawn observation is added to the original data set, leading to $n+1$ observations. This create a partition consisting of $n+2$ intervals, from which the second observation is sampled. This process continues until m observations have been drawn, and these m observations form one NPI-B sample. All possible orderings of the m future values among the n original data observations are equally likely to appear in NPI-B [18, 32, 33].

Assume that the ordered observations of $X_1, \ldots, X_n$ are denoted by $x_1 < x_2 < \ldots < x_n$, and x_0 and x_{n+1} are the end points of the possible data range. The NPI-B method for one-dimensional real-valued data is as follows [18]:

Algorithm 1 NPI-B algorithm

1: The original n observations create $n+1$ intervals $I_j = (x_{j-1}, x_j)$, for $j = 1, ..., n+1$.
2: Sample one of these intervals I_j, where each I_j has probability $\frac{1}{n+1}$.
3: Sample one future value uniformly from this selected interval, and add it to the data set; increase n to $n+1$.
4: Repeat Steps 2 and 3 in total m times to form an NPI-B sample of size m.
5: Repeat Steps 2-4 to create in total B NPI-B samples .

In this NPI-B algorithm, particular attention should be given to Step 3. If the chosen interval in Step 2, $I_1 = (x_0, x_1)$ or $I_{n+1} = (x_n, x_{n+1})$, is a bounded interval, then sample one future value from this interval in a similar way as the other intervals $I_j = (x_{j-1}, x_j), j = 2, \ldots, n$. On the other hand, if the chosen interval is I_1 or I_{n+1} with $x_0 = -\infty$ or $x_{n+1} = \infty$, we sample one future value with probability $\frac{1}{n+1}$ from the interval (x_0, x_1) or (x_n, x_{n+1}) by assuming Normal distribution tail with mean $\mu = \frac{x_1+x_n}{2}$ and standard deviations $\sigma = \frac{x_n-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ^{-1} is the inverse of the normal cumulative distribution function. For the case with $x_0 = 0$ and $x_{n+1} = \infty$, if the chosen interval is I_1, one future value is uniformly sampled as presented in Step 3. If the chosen interval is I_{n+1}, we sample one future value by assuming an Exponential distribution tail with $\lambda = \frac{\ln(n+1)}{x_n}$ [18, 32].

A limitation of the exact NPI-RP method is that the computation of NPI-RP is complicated for large samples due to the increase in the number of orderings of future observations among the

data observations. For some tests deriving the NPI lower and upper reproducibility probabilities is computationally challenging, particularly for parametric tests. This complexity arises because such tests typically involve estimating two parameters, the mean and the variance. To overcome these difficulties, NPI-B can be applied as an alternative method to approximate the NPI reproducibility probability. The NPI-B provides a point estimate for the NPI-RP rather than lower and upper reproducibility probabilities.

In order to derive approximation of the NPI reproducibility probability for statistical tests with $k \geq 2$ groups, the NPI-B method is used to obtain B NPI-B samples per group. Then for run i ($i = 1, 2, \ldots, T$) obtain the proportion of times in which the original data set and the B NPI-B samples lead to the same conclusion, i.e. whether H_0 is rejected or not. Let us denote this proportion as RP_i, $i = 1, 2, \ldots, T$. Then the mean of these RP_i values is the NPI-RP-B estimate for the reproducibility probability (RP) [18, 33]. Other summary statistics (e.g. minimum, median and maximum) based on these RP_i values can also be calculated and used for other inferences.

Algorithm 2 presents NPI-B based approach for estimating the reproducibility probability for statistical hypothesis tests. Since the NPI-B approach is flexible, it will be utilized to estimate the NPI reproducibility probabilities for statistical tests, where it is computationally challenging to derive the exact NPI lower and upper reproducibility probabilities [18].

Algorithm 2 NPI-RP-B algorithm for reproducibility probability for statistical tests

1: Apply the statistical test to the original k independent groups data set, and record the test outcome, that is whether H_0 is rejected or not.
2: Based on the original k-group data set, draw an NPI-B sample from each group with $n = m$, then apply the statistical test on these NPI-B samples.
3: Perform Step 2 B times and each time record the test outcome, whether H_0 is rejected or not.
4: Calculate the proportion of times in which the original k-group data set and the B NPI-B samples lead to the same conclusion, denote that as RP.
5: Perform Steps 2 to 4 in total T times to get RP_i, $i = 1, 2, \ldots, T$. The mean of these values is the NPI-RP-B estimate for the reproducibility probability.

The literature provides suggestions and guidance regarding the question how large the number of bootstrap replications B should be? In the context of estimating a standard error, Tibshirani and Efron [103] suggested that the number of bootstrap replications B is usually between 25 and 200. Even a relatively small number of bootstrap replications, such as $B = 25$ is

usually informative. Tibshirani and Efron [103] discussion indicate that $B = 50$ is often enough to give a good estimate of a standard error. However, it is very seldom that more than $B = 200$ replications are needed for estimating a standard error. To construct confidence intervals and hypothesis tests $B = 100$ or $B = 200$ bootstrap replications is not adequate. Larger values of B, such as $B = 1000$ or $B = 2000$ bootstrap replications are required for obtaining reliable results [103]. The choice of the number of replications B is influenced by many factors, such as the computation time on the computer and the level of precision. In this thesis, $B = 1000$ is considered, which is a common rule of thump in practice and a suitable choice for our objectives.

1.6 Outline of thesis

In this thesis, NPI for reproducibility probability will be investigated for One-way layout tests, using the exact NPI-RP method, the NPI bootstrap and the sampling of orderings methods. This thesis is structured as follows. Chapter 2 introduces NPI reproducibility for the general alternatives tests, namely the Kruskal Wallis (KW) test and the One-way analysis of variance (ANOVA) test. NPI reproducibility for the Jonckheere–Terpstra (JT) test for ordered alternatives is investigated in Chapter 3. Chapter 4 presents NPI reproducibility for umbrella alternatives tests, including the Mack-Wolfe (MW) test and the Esra-Fikri test. In Chapter 5, NPI reproducibility for the Mosteller test for slippage problem is explored. Chapter 6 provides some concluding remarks. Calculations have been done using R [91], and the R codes are available from the author upon request.

Some parts of this thesis was presented in several conferences. The results in Chapters 2 and 3 were presented at the 16th UNCG Regional Mathematics and Statistics Conference in November 2020. Chapters 3 and 4 were presented at the International Conference on Advances in Interdisciplinary Statistics and Combinatorics in October 2021. A comprehensive overview of the main parts of this thesis was presented at the Royal Statistical Society International Conference in September 2022.

Chapter 2

Reproducibility of General Alternatives Tests

2.1 Introduction

Section 1.5 introduced the concept of NPI reproducibility probability. The NPI reproducibility probability is the probability that the same test conclusion would be reached if the test is repeated under similar conditions. This chapter contributes to statistical test reproducibility by considering NPI reproducibility probability for two general alternatives tests, these tests are the parametric One-way analysis of variance (ANOVA) test and the Kruskal-Wallis (KW) test which is the nonparametric analogue of the ANOVA test. The ANOVA test and KW test are statistical hypothesis tests commonly used to determine if there is a difference in the location parameters for three or more independent groups.

As explained in Section 1.5, NPI approach for reproducibility probability involves deriving the exact lower and upper reproducibility probabilities. However, computational issues prevent computing the minimum and maximum values of the ANOVA test statistic and KW test statistic. BinHimd [18] encountered the same computational challenges while exploring the NPI reproducibility probability for the Kolmogorov Smirnov test. She addressed the issue by applying the NPI bootstrap method, introduced in Section 1.5.3, to compute an approximate NPI reproducibility probability. Hence, this chapter focuses on calculating estimates for NPI reproducibility probabilities using NPI bootstrap, which uses the point estimate for the NPI reproducibility probability instead of lower and upper reproducibility probabilities.

In Section 2.2, an overview of the general alternative tests, namely, the ANOVA test and the KW test is provided. In Section 2.3, the NPI-RP-B approach is introduced to study the reproducibility for the ANOVA test and the KW test. Section 2.4 presents application examples

with data sets from the literature, where the NPI-RP-B approach is considered. In Section 2.5, the reproducibility of the ANOVA test and the KW test is investigated using the NPI-RP-B approach, via simulations under both the null and the alternative hypothesis. We conclude the content of this chapter in Section 2.6.

2.2 General alternative tests

Hypothesis testing in practice often involves comparing groups of observations and testing general rather than specific differences among groups means. For example, an investigator might be interested in the sources of variation in patients' blood cholesterol level when using three different drug formulations. General alternative tests is a core technique for analysing such information. General alternatives tests are statistical hypothesis tests used to determine if there are statistically significant differences between the location parameters of $k \geq 3$ groups, that is at least one location parameter is different. The null and alternative hypotheses are as follows:

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{2.1}$$

$$H_1 : \text{at least one } \mu_i \text{ is different.} \tag{2.2}$$

where μ_i is the location parameter of the ith group. In this chapter, we consider the reproducibility of the ANOVA test and the KW test.

2.2.1 Assumptions for general alternatives tests

Most statistical tests are generally based on a number of assumptions. It is important to assess the assumptions before proceeding with any relevant statistical procedure for reliable conclusions. In this section, we review the assumptions underlying the ANOVA test and the KW test. The ANOVA test is based on three assumptions: the assumption of normality where it is required that the observations are drawn from normally distributed populations, the homogeneity of variance assumption in which the variances of all populations are equal and the independence of observations. The dependence between observations can be avoided by selecting the appropriate sampling methods in collecting data. The assumption of normality can be tested statistically using Shapiro–Wilk test, or checked graphically using plots such as histograms, quantile-quantile plots. The Shapiro-Wilk test is one of the most common normality test procedures available in statistics. The test was proposed by Shapiro and Wilk [96], to detect departures from normality. There are other tests to asses normality, such as the Kolmogorov-Smirnov test, Anderson-Darling test, but the Shapiro-Wilk test provides better power than the

other tests [102]. The homogeneity of variance assumption can be tested by Bartlett's test or by visualization, such as box plots for each group and by plotting the residuals to find out whether they have a similar spread [46]. However, Razali and Wah [92] argue that graphical methods can be subjective and may not provide conclusive evidence that the assumptions hold. Therefore, to support the graphical methods, formal tests can be performed.

There are situations in which the ANOVA test assumptions are violated. Such violations can seriously affect the validity of the statistical conclusions. Nonparametric tests such as the Kruskal-Wallis test should be considered. The assumptions of the Kruskal-Wallis test include independence of observations within and among groups, the variable of interest being continuous and data measured on an ordinal scale. The Kruskal–Wallis test does not make assumptions about normality and have been commonly used to test hypothesis about the location parameters. [15, 55].

2.2.2 One-way Analysis of Variance (ANOVA) test

One-way analysis of variance (ANOVA) is a statistical technique used to compare the means of more than two independent groups. Let us first introduce some notations. The data consist of $N = \sum_{i=1}^{k} n_i$ observations, with n_i observations for the ith group where $i = 1, 2, ..., k$. Let x_{ij} represents the jth observation in the ith group [67]. The mean for the ith group is given by:

$$\overline{x}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$$

The dot indicates the aggregation over the j index. The total of all observations is denoted by $x_{..}$:

$$x_{..} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

where the dots indicate the aggregation over the i and j indexes. The overall mean for all observations is represented by $\overline{x}_{..}$:

$$\overline{x}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} x_{ij}$$

The total sum of squares of the observations about the overall mean can be partitioned as follows:

$$\sum_i \sum_j (x_{ij} - \bar{x_{..}})^2 = \sum_i \sum_j (x_{ij} - \bar{x_{i.}})^2 + \sum_i n_i (\bar{x_{i.}} - \bar{x_{..}})^2 \tag{2.3}$$

Thus, Equation (2.3) can be expressed symbolically as

$$SST = SSE + SSR \tag{2.4}$$

where SSE is the sum of squares due to error within groups, and SSR is the sum of squares between groups. ANOVA uses the F-test for testing the equality of groups means [63], as follows

$$F = \frac{\frac{SSR}{k-1}}{\frac{SSE}{N-k}} \tag{2.5}$$

The null hypothesis is rejected if $F > F(1-\alpha; k-1, N-k)$. The effect size estimate, η^2, can be computed as follows

$$\eta^2 = \frac{SSR}{SST} \tag{2.6}$$

where the index η^2 assumes values from 0 to 1, and measures the proportion of total variance that can be explained by the differences between groups [23, 60, 104]. For the interpretation of the effect size strength, Cohen [24] recommends the rule of thumb: $0.10 \leq \eta^2 <$0.25: small effect, $0.25 \leq \eta^2 <$ 0.40: medium effect, $0.40 \leq \eta^2$: large effect.

2.2.3 Kruskal–Wallis (KW) test

Nonparametric tests are statistical procedures that do not rely on the assumption that the data are drawn from a parametric family of probability distributions. The nonparametric equivalent to the ANOVA test is the Kruskal–Wallis (KW) test. The Kruskal–Wallis test is used for testing the equality of the probability distributions for more than two independent groups when the assumptions of the parametric ANOVA test are not met.

To compute the Kruskal–Wallis test statistic, KW, we first combine all N observations from the k groups and rank them from smallest to largest values (giving each observation in a group of ties the mean of the ranks tied). Let r_{ij} denote the rank of x_{ij} in this joint ranking, $i = 1, \ldots, k$ and $j = 1, 2, \ldots, n_i$. Let $R_i = \sum_{j=1}^{n_i} r_{ij}$ be the sum of ranks of observations from the ith group and let $\bar{R}_i = \frac{R_i}{n_i}$. Thus, for example, R_1 is the sum of the joint ranks received by the first group observations and $\bar{R}_1$ is the average rank for the first group observations [59]. The Kruskal–Wallis test statistic, KW, is defined as:

$$KW = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \Big(\bar{R}_i - \frac{N+1}{2}\Big)^2$$

where $\sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{r_{ij}}{N} = \frac{N+1}{2}$ is the mean rank assigned in the joint ranking, and $E(R_i) = \frac{N+1}{2}$ under H_0 [59]. The null hypothesis is rejected, at level of significant α, if and only if

$$KW \geq h_\alpha \tag{2.7}$$

with critical value h_α which can be found from tables in [50, 59]. Under the null hypothesis, the statistic KW has, as $n_i \to \infty$, an asymptotic Chi-square distribution with $k-1$ degrees of

freedom [66]. For the chi-square approximation, H_0 is rejected if and only if

$$KW \geq \chi^2_{k-1,\alpha} \tag{2.8}$$

where $\chi^2_{k-1,\alpha}$ is the upper $100(1-\alpha)$ percentile of a Chi-square distribution with $k-1$ degree of freedom. The effect size for the Kruskal-Wallis test, ε^2, can be calculated as follows [64],

$$\varepsilon^2 = \frac{KW}{(N^2-1)/(N+1)} \tag{2.9}$$

Using the R command `Kruskal.test` one can get the value of the test statistic and the p-value. The exact and asymptotic critical values can be obtained using the commands $cKW(\alpha, c(n1, ..., n_i)$, "*Exact*") `and` $cKW(\alpha, c(n1, \ldots, n_i)$`,"Asymptotic")` in the R package `NSM3` with R version 4.2.2.

2.3 NPI-RP-B for the general alternatives tests

This section investigates the reproducibility probability for the KW test and the ANOVA test, using the NPI-RP-B method from Section 1.5.3, with the implementation of Algorithm 2. This Algorithm uses NPI bootstrap to derive reproducibility probability for a statistical test. The inputs into Algorithm 2 are the k original samples, their corresponding sample sizes, the number of runs T and the number of bootstrapped samples per run B. Summary statistics: (e.g., minimum, mean, median and maximum) of $RP_1, RP_2, \ldots, RP_T$ were calculated, where the mean of $RP_1, RP_2, \ldots, RP_T$ is the reproducibility probability estimate, and is referred to as NPI-RP-B value. In Section 2.4, the selection of values of T and B is explored. The NPI-RP-B approach is considered in this thesis with finite and infinite intervals, labeled as Approach I and Approach II, respectively. Let the order of the observations $X_1, \ldots, X_n$ be denoted by $x_1 < x_2 < \ldots < x_n$. The n ordered observations creates $n+1$ intervals: $((x_0, x_1), (x_1, x_2), \ldots, (x_{n-1}, x_n), (x_n, x_{n+1}))$, where x_0 and x_{n+1} are the end points of the possible data range.

(I) $x_0 = x_1 - \max(x_j - x_{j-1})$ and $x_{n+1} = x_n + \max(x_j - x_{j-1})$, where $j = 1, \ldots, n$.

(II) For the case with data on the real line $(-\infty, \infty)$: $x_0 = -\infty$ and $x_{n+1} = \infty$. If the chosen interval $(-\infty, x_1)$ or (x_n, ∞), Normal distribution tails are assumed with mean $\mu = \frac{x_1+x_n}{2}$ and standard deviations $\sigma = \frac{x_n-\mu}{\Phi^{-1}(\frac{n}{n+1})}$, where Φ^{-1} is the inverse of the normal cumulative distribution function. For the case with date on $[0, \infty)$: $x_0 = 0$ and $x_{n+1} = \infty$. If the chosen interval is (x_n, ∞), one future value is sampled from this interval by assuming an Exponential distribution tail with $\lambda = \frac{\ln(n+1)}{x_n}$.

Manufacturing	18.79	20.22	20.25	22.46	24.74	27.97	28.19
	28.66	29.18	29.52	31.64	31.99		
Marketing	23.01	27.63	29.361	29.92	31.06	31.22	33.18
	33.41	35.22	35.33	36.50	37.03	37.81	37.89
Research	25.44	25.70	26.28	26.54	26.64	27.12	28.90
	31.90	32.05	33.42	35.78			

Table 2.1: The employees satisfaction scores, for Example 2.1

Throughout this thesis, the results in the tables were rounded to three decimal digits and precise value 1 is presented without additional decimals, so the values 1.000 are less than 1 but rounded up. Furthermore, the test outcome is either to reject (R) or to not reject (NR) the null hypothesis. In Section 2.4, data sets from the literature will be used to investigate reproducibility for the KW test and the ANOVA test. In Section 2.5, the results of simulation studies for different scenarios are presented, such as simulation under H_0 and under H_1, with varying sample sizes and number of groups. Section 2.5 also studies the relationship between the p-value and the NPI reproducibility probability, and between the effect size and the NPI reproducibility probability for both tests. Reproducibility of the KW test and the ANOVA test is also briefly compared.

2.4 Examples

This section investigates the reproducibility probability for the KW test and the ANOVA test, using data sets from the literature. In Example 2.1, we will examine how the choice of T and B affects the computational time and the accuracy of the reproducibility probability estimates. In Example 2.2, we will compare the reproducibility probability results of the finite and infinite intervals.

Example 2.1. The data set, in Table 2.1, originates from a study conducted by a company to assess employees reactions to a newly implemented salary and fringe benefits plan. Random samples of 15 employees were taken from each of three divisions: manufacturing, marketing, and research. The personnel staff asked each employee sampled to respond to a series of questions. Several employees refused to cooperate, as reflected in the unequal sample sizes. The average responses from the employees are provided, with larger scores reflecting a higher degree of satisfaction with management [85]. The satisfaction scores from the three divisions: manufacturing, marketing, and research, are labeled as X, Y, and Z, respectively. Figure 2.1

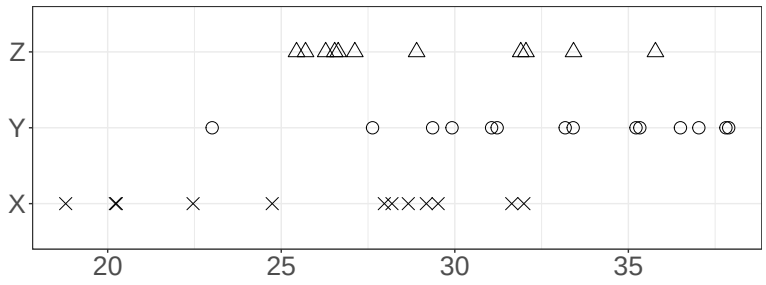


Figure 2.1: Visualization of the employees satisfaction scores data, for Example 2.1

provides visualization of the data. The size of the groups are $n_x = 12$, $n_y = 14$, $n_z = 11$. For groups Y and Z, the observation 33.41 is a tied observation; we break this tied observation by adding a small amount to make it 33.42 in group Z, as explained in Section 1.4. To test the hypothesis $H_0 : \mu_x = \mu_y = \mu_z$ against H_1 : at least one μ_i is different, the level of significance is set at $\alpha = 0.05$.

Before applying the ANOVA test, assessments were carried out to check its underlying assumptions. The Shapiro-Wilk test suggests that the normally assumption is satisfied for the groups X, Y and Z, yielding p-values of 0.153, 0.356 and 0.068, respectively. Furthermore, the Bartlett test of homogeneity of variances yields a non-significant p-value of 0.713, indicating no significant evidence of unequal variances. Since the assumptions are met, the ANOVA test is applied to the data and the original test outcome is obtained. The p-values for the KW test and the ANOVA test are 0.005 and 0.001, respectively. So, the null hypothesis is rejected for both tests at the significance level $\alpha = 0.05$. This indicates that there is evidence of significant differences in the mean satisfaction scores for the three company divisions.

In Table 2.2, Algorithm 2 is implemented using different values for the number of runs T and the number of bootstrapped samples per run B, with infinite support (Approach II). The reproducibility probability is considered with $B = 1000, 10000$ and $T = 100, 200, 500$. The results in Table 2.2 show that increasing T from 100 to 200 or to 500 slightly expand the range between the minimum and maximum of $RP_1, RP_2, \ldots, RP_T$. However, the change is minor with the mean value differing only in the third decimal place. Increasing B from 1,000 to 10,000, the mean and the median differed only in the third decimal. Increasing B and T leads to larger computational time by about the same amount, without any notable increase in the accuracy of the RP estimates. Therefore, in this thesis, the choice of B will be set at 1000 and the choice of T with be set at 100. This selection of B and T achieves a practical balance between the accuracy of the RP estimates and the computational time. Moreover, the RP estimates in Table 2.2 are typically large due to p-values not close to the threshold 0.05. The comparison of the reproducibility results for the KW test and the ANOVA test indicates that they are quite

Replications		KW				ANOVA			
B	T	Min	Mean	Median	Max	Min	Mean	Median	Max
1000	100	0.751	0.788	0.789	0.825	0.772	0.803	0.806	0.837
1000	200	0.749	0.788	0.789	0.825	0.771	0.803	0.804	0.837
1000	500	0.749	0.788	0.788	0.825	0.771	0.802	0.802	0.847
10,000	100	0.777	0.788	0.788	0.797	0.792	0.803	0.803	0.812
10,000	200	0.774	0.788	0.788	0.797	0.792	0.803	0.803	0.812
10,000	500	0.774	0.788	0.788	0.799	0.792	0.802	0.803	0.812

Table 2.2: RP for the KW test and the ANOVA test, for Example 2.1

X	38.7	41.5	43.8	44.8	45.5
Y	39.2	39.3	39.7	41.4	41.8
Z	34.0	35.0	39.0	40.0	43.0
V	34.1	34.8	34.9	35.4	37.2

Table 2.3: Smoothness of papers data, for Example 2.2

similar.

Example 2.2. This example is introduced to study the NPI-RP for the KW test and the ANOVA test using the data given in Table 2.3, and visualized in Figure 2.2 [50], where there are four sets X, Y, Z and V of five measurements of the smoothness of a certain type of papers, each set is obtained from different laboratory. The aim is to test if the smoothness is the same for all laboratories.

Algorithm 2 has been applied with finite range (Approach I) and infinite range (Approach II), to investigate the impact of range on the NPI-RP for the KW test and ANOVA test. To test the hypothesis H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ against H_1: at least one μ_i is different, the level of significance is set at $\alpha = 0.05$. In Table 2.4, Case 1 refers to the case where three groups of smoothness of papers X, Y, Z are considered to study NPI-RP, with $n = 5$. Case 2 refers to the case where all four groups of smoothness of a paper X, Y, Z, V are considered, with $n = 5$.

In order to apply the ANOVA test, tests were performed to check the assumptions for both cases. The Shapiro-Wilk test for groups X, Y, Z and V results in p-values 0.520, 0.136, 0.687 and 0.354, respectively, indicating that the normally assumption is met. The Bartlett test of homogeneity of variances yields p-values of 0.157 and 0.085 for Case 1 and 2, respectively. Therefore, there is no significant evidence of unequal variances. For Case 1, the KW test and

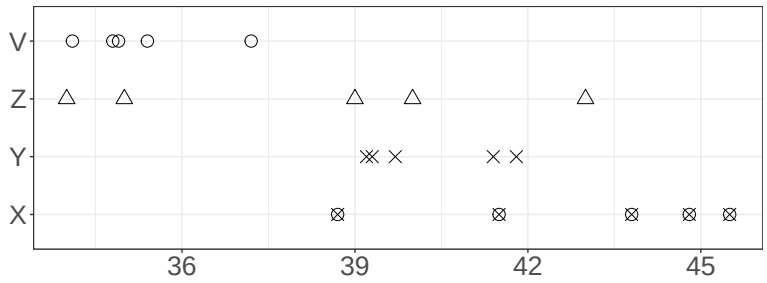


Figure 2.2: Visualization of the smoothness of papers data, for Example 2.2

Cases	NPI-B	KW				ANOVA			
		Min	Mean	Median	Max	Min	Mean	Median	Max
Case 1	Approach(I)	0.534	0.570	0.570	0.619	0.482	0.524	0.525	0.575
	Approach(II)	0.537	0.575	0.578	0.608	0.502	0.536	0.537	0.565
Case 2	Approach(I)	0.863	0.890	0.889	0.910	0.868	0.898	0.898	0.916
	Approach(II)	0.839	0.856	0.856	0.879	0.829	0.860	0.859	0.885

Table 2.4: RP for the KW test and ANOVA test, for Example 2.2

the ANOVA test p-values are 0.137 and 0.061, respectively. Thus, the null hypothesis is not rejected for both tests in Case 1. For Case 2, the KW test and the ANOVA test p-values are 0.014 and 0.001. So, considering group V in Case 2 leads to the null hypothesis being rejected for both tests.

Based on the NPI-RP estimates in Table 2.4, for Case 1, we have less trust in the decision that we are going to get the same conclusion in the future tests. As can be seen in Table 2.4, the reproducibility does not change notably when using different ranges for bootstrapping in Algorithm 2. For Case 2, the reproducibility is relatively large due to the p-values not close to the threshold $\alpha = 0.05$. It can be inferred from the comparison of the reproducibility results for the KW test and the ANOVA test that the reproducibility for both tests are quite similar.

2.5 Simulation study

This section studies reproducibility probability for the KW test and the ANOVA test via simulations, where reproducibility is calculated using Algorithm 2. The NPI-RP-B approach in Section 2.3 is considered using Approach II, which involves using the tail of a Normal distribution for real-valued data, and the tail of an Exponential distribution for non-negative real-valued data. The null hypothesis is $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$ against H_1 : at least oneμ_i is different, the level of significance is $\alpha = 0.05$. Data were simulated under H_0 and H_1, as presented in Table

Case	k	Simulation
1	3	$N(0,1)$
2	3	Gamma(2, 1)
3	3	$X \sim N(0,1)$, $Y \sim N(0,1)$, $Z \sim N(1.5,1)$
4	5	$N(0,1)$

Table 2.5: Simulation cases for the KW test and the ANOVA test

2.5. To study the impact of the number of groups and the sample size on the reproducibility probability, the simulation is considered with the number of groups $k = 3, 5$ and the sample size $n = 6, 10, 20$. Under H_0, the original data were generated from the Normal distribution with mean 0 and standard deviation 1 for $k = 3$ and $k = 5$ groups. Under H_0 and with $k = 3$ groups, the original data were generated from the Gamma distribution with shape parameter 2 and scale parameter 1. Under H_1 and with $k = 3$ groups, data were generated from Normal distribution with different means $\mu_x = \mu_y = 0$ and $\mu_z = 1.5$ and the standard deviation 1. Further simulations were performed for data generated under H_1 for $k = 5$ and with sample sizes $n = 6, 10, 20$, the results are presented in the Appendix A.

The inputs for the simulation study in Tables 2.6 through 2.17 are as follows: Algorithm 2 is applied with $B = 1000$ and $T = 100$. For each run, one sample of size n is generated from each of the distributions given in Table 2.5, the KW test and the ANOVA test are performed both on these samples, and the tests outcomes are obtained and the RP estimates for the KW test and the ANOVA test are calculated using Algorithm 2. In each Table, the reproducibility probability estimates have been reported for 10 simulated original data sets. Effect size, introduced in Sections 2.2.2 and 2.2.3, has also been calculated for the tests. Note that the threshold values, introduced in Sections 2.2.2 and 2.2.3, are provided in the caption of each table for both tests. As expected, the worst reproducibility probability occurs when the p-value is close to the threshold 0.05, regardless of the decision about H_0. The reproducibility probability starts to increase when moving away from the threshold leading to high estimates of the reproducibility probability. Similar patterns have been observed in previous applications of NPI studies investigating tests reproducibility [29, 32, 33, 75, 98]. For the same p-value or the same effect size value, reproducibility probability estimates differs from one data set to another. These small variations in the RP estimates are due to variations in the original samples and in the NPI-B samples.

The relationship between NPI-RP-B and the p-value for the KW test and the ANOVA test is examined in the simulations. The minimum and maximum of the RP values are also added

to the plots. We use the p-value for better visualization of figures rather than the critical value because each simulation scenario has a different critical value, given the variations in the sample sizes and the number of groups considered. Although the p-values and critical values are two different approaches, they ultimately yield the same conclusion regarding whether the null hypothesis is rejected or not. For simulations under H_0 in Figures 2.3, 2.4 and 2.6, there are few cases where the null hypothesis is wrongly rejected which aligns with where the test is set up. Note that the level of significance $\alpha = 0.05$ is represented on the figures by a vertical line. For simulations under H_1 in Figure 2.5, since we sample from a case which is in line with the alternative hypothesis, there are more cases where the null hypothesis is rejected. Increasing the size of samples leads to increasing the power of the test and more cases rejecting the null hypothesis, as for sample sizes $n = 10$ and $n = 20$ in Figure 2.5. Consequently, there is a tendency for RP estimates to be higher in cases of rejection than in non-rejection, as shown in Figure 2.6 compared to Figure 2.3. The reason is that in the case of rejection we obtain more cases of the same decision of an original sample that does reject H_0. The values of RP tend to increase with increasing distance between the observed p-value and the threshold $\alpha = 0.05$, regardless of the decision about H_0. In this section, we consider increasing n up to 20, however, considering larger n will lead to relatively lower reproducibility in non-rejection cases compared to small n. One noteworthy observation becomes apparent: as n increases, specifically to $n = 20$, the reproducibility probability curve becomes progressively smoother. Figure 2.4 represents Case 2, where we generated data from a scenario in which the normality assumption of ANOVA test is violated, hence there is some variability in the RP results for ANOVA test, and the variability remains relatively constant as n increases compared to KW test where the variability decreases as n increases. The reason is that the ANOVA test assumes that the data are normally distributed and violations of this assumption impact the RP results. However, when the data were generated from Case 1, 3 and 4 where the ANOVA assumptions are met, as shown in Figures 2.3, 2.5 and 2.6, there is low variability in the reproducibility probability results for the ANOVA test compared to the KW test.

The relationship between NPI-RP-B and the effect size for the KW test and the ANOVA test is studied. Figures 2.7 and 2.8 show the results of simulations under H_0 and under H_1, where Case 1 and Case 3, introduced in Table 2.5 are considered, respectively. In Figures 2.7 and 2.8, there is a V-shape pattern and the NPI-RP-B estimates tend to increase as the effect size moves away from the area where the V-shape has the lowest point. It is also observed that, as the sample size n increases, the location of the lowest point shifts to the left and the range of effect size in the figures tends to be smaller. There is a tendency for NPI-RP-B estimates to

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
8.187	0.017	R	0.482	0.588	0.623	0.622	0.655	5.525	0.016	R	0.420	0.617	0.645	0.645	0.679
6.035	0.049	R	0.355	0.538	0.567	0.568	0.603	4.689	0.026	R	0.380	0.583	0.620	0.617	0.661
5.415	0.067	NR	0.319	0.511	0.550	0.552	0.582	2.103	0.157	NR	0.220	0.555	0.598	0.599	0.631
5.099	0.078	NR	0.300	0.522	0.559	0.560	0.593	3.213	0.069	NR	0.300	0.487	0.535	0.534	0.574
5.099	0.078	NR	0.300	0.502	0.538	0.536	0.584	3.062	0.077	NR	0.290	0.489	0.529	0.528	0.578
4.538	0.103	NR	0.267	0.502	0.542	0.540	0.584	3.379	0.061	NR	0.311	0.439	0.490	0.487	0.534
3.135	0.209	NR	0.184	0.637	0.669	0.667	0.703	1.821	0.196	NR	0.195	0.612	0.644	0.646	0.677
2.608	0.271	NR	0.153	0.678	0.711	0.711	0.739	0.902	0.427	NR	0.107	0.688	0.722	0.721	0.757
1.731	0.421	NR	0.102	0.702	0.735	0.736	0.770	0.879	0.436	NR	0.105	0.685	0.718	0.719	0.753
0.035	0.983	NR	0.002	0.779	0.817	0.819	0.848	0.078	0.926	NR	0.010	0.769	0.796	0.796	0.822

Table 2.6: RP under H_0, with Case 1, $n = 6$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 15) = 3.682$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
11.848	0.003	R	0.409	0.795	0.824	0.824	0.854	8.707	0.001	R	0.392	0.814	0.842	0.843	0.871
6.521	0.038	R	0.225	0.558	0.595	0.596	0.636	3.097	0.062	NR	0.187	0.433	0.465	0.465	0.496
5.515	0.063	NR	0.190	0.441	0.479	0.478	0.513	3.251	0.054	NR	0.194	0.419	0.463	0.462	0.509
4.978	0.083	NR	0.172	0.444	0.484	0.485	0.516	2.912	0.072	NR	0.177	0.436	0.484	0.483	0.518
3.440	0.179	NR	0.119	0.541	0.585	0.586	0.628	1.947	0.162	NR	0.126	0.526	0.574	0.577	0.611
3.223	0.200	NR	0.111	0.567	0.612	0.615	0.647	1.130	0.338	NR	0.077	0.630	0.664	0.665	0.692
2.728	0.256	NR	0.094	0.567	0.619	0.617	0.655	1.870	0.174	NR	0.122	0.548	0.594	0.592	0.630
2.728	0.256	NR	0.094	0.622	0.655	0.655	0.687	1.223	0.310	NR	0.083	0.626	0.663	0.664	0.694
1.030	0.598	NR	0.036	0.709	0.749	0.750	0.787	0.335	0.718	NR	0.024	0.731	0.762	0.761	0.795
0.034	0.983	NR	0.001	0.758	0.801	0.801	0.835	0.102	0.903	NR	0.007	0.746	0.787	0.788	0.819

Table 2.7: RP under H_0, with Case 1, $n = 10$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 27) = 3.354$

be higher in cases of rejection than in non-rejection. Further, for the non-rejection cases, there is a linear relationship between the effect size and the NPI-RP-B estimates.

To sum up, the comparison of the reproducibility for the KW test and the ANOVA test shows similar patterns in the reproducibility probability results across the different distribution parameters and sample sizes. Specifically, the NPI-RP estimates for both tests generally tend to increase as the test statistic moves away from the test thresholds, regardless of the decision about H_0.

The time taken to run the R code for each simulated data set using Algorithm 2, varied with the number of groups and sample sizes. For three groups with a sample size of 6, the time is 1 minute and 39 seconds. Increasing the sample size to 10 extended the runtime to 1 minute and 59 seconds. With a further increase in the sample size to 20, the runtime is 3 minutes. For five groups with a sample size of 6, the runtime is approximately 2 minutes and 22 seconds.

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
9.031	0.011	R	0.153	0.682	0.726	0.727	0.764	6.172	0.004	R	0.178	0.738	0.778	0.777	0.812
5.996	0.050	R	0.102	0.535	0.592	0.591	0.637	2.774	0.071	NR	0.089	0.438	0.469	0.469	0.517
5.673	0.059	NR	0.096	0.391	0.430	0.429	0.476	3.144	0.051	NR	0.099	0.374	0.411	0.411	0.464
5.170	0.075	NR	0.088	0.409	0.448	0.448	0.496	3.198	0.048	R	0.101	0.538	0.576	0.575	0.616
4.738	0.094	NR	0.080	0.450	0.503	0.506	0.540	1.494	0.233	NR	0.050	0.567	0.613	0.613	0.653
3.845	0.146	NR	0.065	0.502	0.540	0.541	0.585	1.375	0.261	NR	0.046	0.587	0.623	0.623	0.660
2.047	0.359	NR	0.035	0.593	0.641	0.643	0.674	1.235	0.299	NR	0.042	0.600	0.633	0.634	0.672
1.432	0.489	NR	0.024	0.645	0.687	0.687	0.719	0.711	0.495	NR	0.024	0.664	0.695	0.695	0.726
1.001	0.606	NR	0.017	0.662	0.712	0.714	0.746	1.041	0.360	NR	0.035	0.610	0.655	0.653	0.688
0.042	0.979	NR	0.001	0.738	0.774	0.776	0.813	0.006	0.994	NR	2.019×10^{-4}	0.758	0.788	0.789	0.828

Table 2.8: RP under H_0, with Case 1, $n = 20$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 57) = 3.159$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
5.626	0.060	NR	0.331	0.443	0.492	0.491	0.530	4.135	0.037	R	0.355	0.483	0.524	0.523	0.554
5.135	0.077	NR	0.302	0.522	0.557	0.558	0.599	4.060	0.039	R	0.350	0.434	0.467	0.463	0.504
3.895	0.143	NR	0.229	0.586	0.635	0.635	0.668	2.584	0.109	NR	0.260	0.605	0.641	0.642	0.685
3.696	0.158	NR	0.217	0.582	0.614	0.614	0.646	1.442	0.267	NR	0.161	0.660	0.703	0.704	0.734
2.889	0.236	NR	0.170	0.622	0.666	0.667	0.704	1.720	0.213	NR	0.187	0.641	0.679	0.681	0.708
2.351	0.309	NR	0.138	0.678	0.711	0.711	0.751	1.988	0.172	NR	0.210	0.629	0.669	0.670	0.702
1.977	0.372	NR	0.116	0.689	0.719	0.718	0.763	0.992	0.394	NR	0.117	0.693	0.731	0.731	0.770
1.205	0.548	NR	0.071	0.763	0.793	0.795	0.828	0.134	0.875	NR	0.020	0.756	0.800	0.801	0.834
0.784	0.676	NR	0.046	0.759	0.789	0.789	0.829	0.150	0.862	NR	0.020	0.795	0.829	0.830	0.857
0.082	0.960	NR	0.005	0.785	0.818	0.817	0.852	0.250	0.782	NR	0.030	0.773	0.815	0.815	0.845

Table 2.9: RP under H_0, with Case 2, $n = 6$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 15) = 3.682$

When the sample size is increased to 10, the runtime is 3 minutes. Finally, for five groups with a sample size of 20, the runtime reached 4 minutes and 48 seconds.

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
5.853	0.054	NR	0.202	0.436	0.471	0.471	0.510	2.958	0.069	NR	0.180	0.503	0.537	0.535	0.579
5.267	0.072	NR	0.182	0.445	0.489	0.490	0.528	3.276	0.0532	NR	0.195	0.448	0.479	0.479	0.508
4.222	0.121	NR	0.146	0.506	0.553	0.554	0.596	3.927	0.0319	NR	0.230	0.466	0.505	0.505	0.545
3.177	0.204	NR	0.110	0.572	0.620	0.622	0.660	1.811	0.183	NR	0.118	0.608	0.641	0.642	0.673
2.996	0.224	NR	0.103	0.593	0.632	0.633	0.684	1.407	0.262	NR	0.090	0.638	0.674	0.676	0.708
2.712	0.258	NR	0.094	0.597	0.631	0.631	0.686	1.577	0.225	NR	0.105	0.623	0.668	0.668	0.695
2.023	0.364	NR	0.070	0.641	0.679	0.679	0.709	1.231	0.308	NR	0.080	0.633	0.674	0.674	0.709
1.435	0.488	NR	0.049	0.649	0.690	0.691	0.728	0.978	0.389	NR	0.068	0.662	0.701	0.700	0.732
1.056	0.590	NR	0.036	0.693	0.725	0.725	0.766	1.045	0.365	NR	0.070	0.672	0.706	0.706	0.737
0.436	0.804	NR	0.015	0.753	0.782	0.782	0.812	0.066	0.936	NR	0.005	0.775	0.804	0.805	0.831

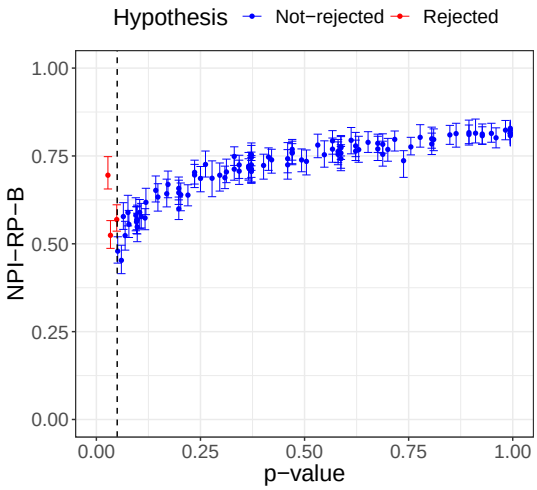
Table 2.10: RP under H_0, with Case 2, $n = 10$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 27) = 3.354$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
5.912	0.052	NR	0.100	0.403	0.443	0.443	0.478	1.643	0.202	NR	0.055	0.578	0.613	0.613	0.646
4.432	0.109	NR	0.075	0.479	0.517	0.516	0.559	2.254	0.114	NR	0.073	0.507	0.549	0.549	0.589
3.848	0.146	NR	0.065	0.495	0.537	0.540	0.585	3.049	0.0552	NR	0.097	0.414	0.456	0.460	0.491
3.779	0.151	NR	0.064	0.504	0.549	0.548	0.582	1.272	0.288	NR	0.040	0.611	0.650	0.651	0.695
3.032	0.220	NR	0.051	0.560	0.591	0.590	0.630	2.243	0.115	NR	0.073	0.529	0.559	0.558	0.603
2.816	0.245	NR	0.048	0.557	0.593	0.593	0.625	1.556	0.22	NR	0.052	0.567	0.605	0.605	0.652
2.492	0.288	NR	0.042	0.557	0.613	0.615	0.645	0.559	0.575	NR	0.019	0.675	0.715	0.717	0.743
1.570	0.456	NR	0.027	0.634	0.672	0.674	0.710	1.199	0.309	NR	0.040	0.633	0.667	0.666	0.720
1.092	0.579	NR	0.019	0.674	0.711	0.710	0.745	0.404	0.670	NR	0.014	0.708	0.742	0.743	0.785
0.439	0.803	NR	0.007	0.706	0.756	0.757	0.786	0.025	0.976	NR	0.001	0.764	0.795	0.796	0.834

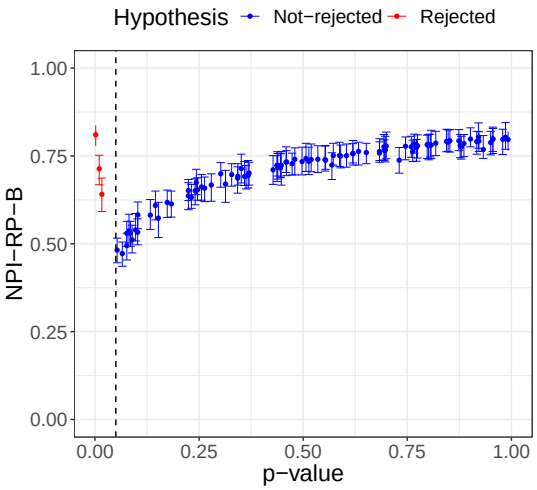
Table 2.11: RP under H_0, with Case 2, $n = 20$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 57) = 3.159$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
12.737	0.002	R	0.749	0.945	0.961	0.961	0.973	22.960	2.722×10^{-5}	R	0.754	0.960	0.972	0.972	0.983
9.977	0.007	R	0.587	0.794	0.831	0.831	0.864	9.245	0.002	R	0.552	0.809	0.842	0.842	0.869
7.614	0.022	R	0.448	0.739	0.773	0.773	0.801	6.024	0.012	R	0.445	0.723	0.766	0.766	0.794
6.351	0.042	R	0.374	0.485	0.515	0.517	0.563	4.393	0.032	R	0.369	0.525	0.555	0.556	0.597
6.351	0.042	R	0.374	0.539	0.565	0.564	0.597	4.670	0.027	R	0.384	0.543	0.584	0.583	0.623
5.661	0.059	NR	0.333	0.460	0.509	0.509	0.540	4.016	0.040	R	0.349	0.491	0.524	0.525	0.574
4.924	0.085	NR	0.290	0.531	0.574	0.574	0.610	3.247	0.067	NR	0.302	0.497	0.544	0.544	0.582
3.790	0.150	NR	0.223	0.599	0.643	0.643	0.673	2.266	0.138	NR	0.232	0.573	0.609	0.610	0.644
1.825	0.402	NR	0.107	0.693	0.728	0.728	0.756	1.069	0.368	NR	0.125	0.655	0.695	0.695	0.738
0.327	0.849	NR	0.019	0.770	0.799	0.798	0.829	0.205	0.817	NR	0.027	0.752	0.786	0.788	0.827

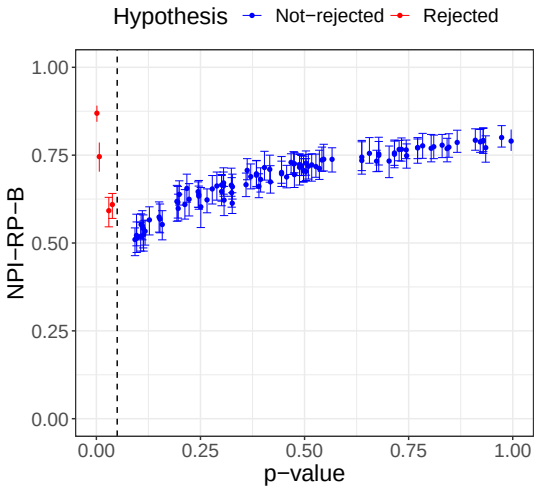
Table 2.12: RP under H_1, with Case 3, $n = 6$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 15) = 3.682$



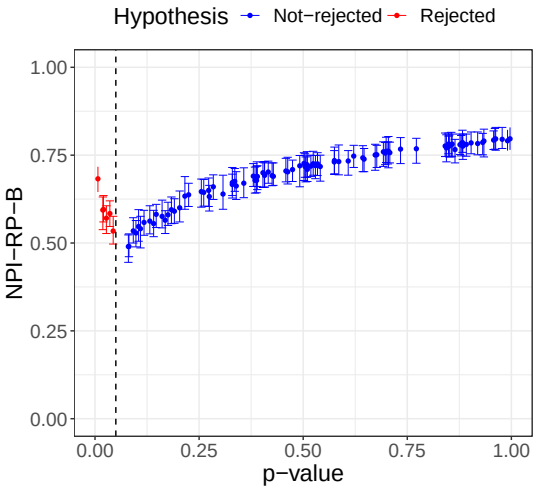
(a) KW, $n = 6$



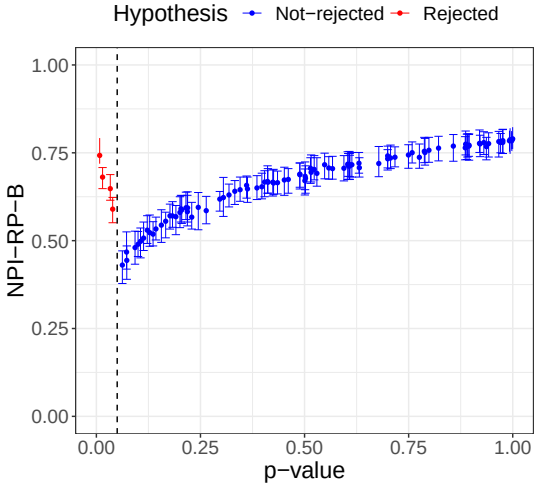
(b) ANOVA, $n = 6$



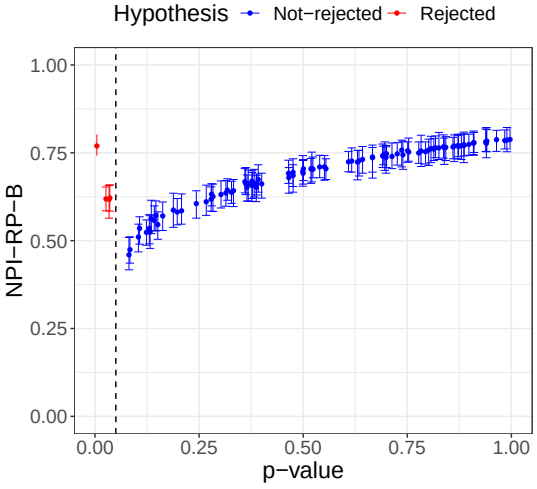
(c) KW, $n = 10$



(d) ANOVA, $n = 10$

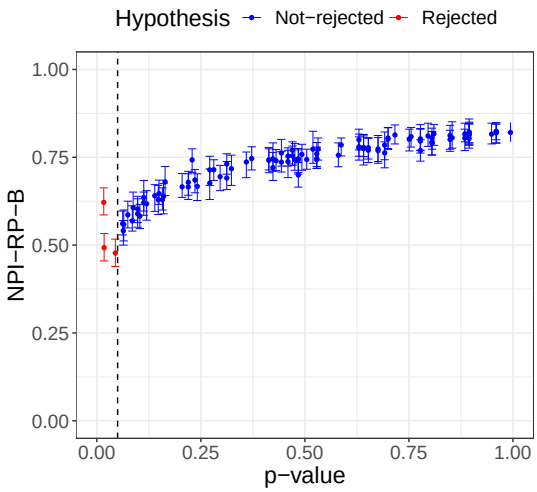


(e) KW, $n = 20$

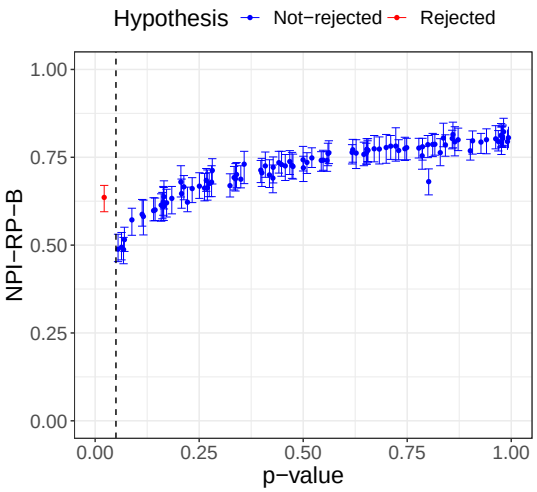


(f) ANOVA, $n = 20$

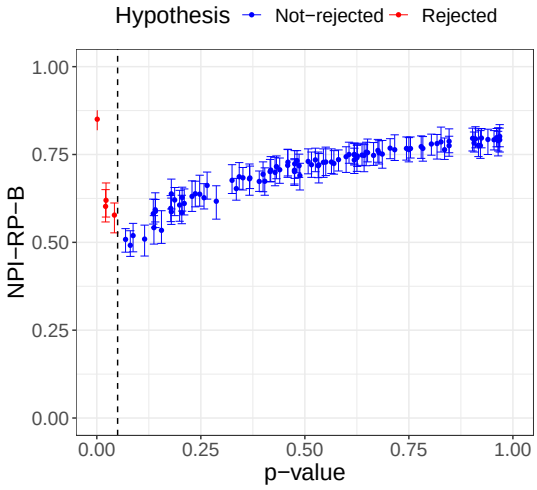
Figure 2.3: NPI-RP-B under H_0, with Case 1, $\alpha = 0.05$



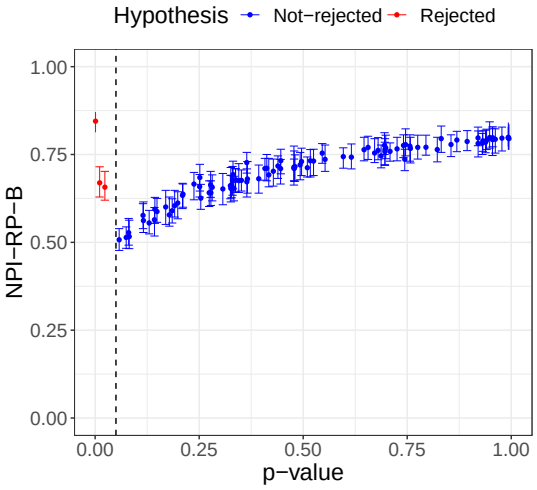
(a) KW, $n = 6$



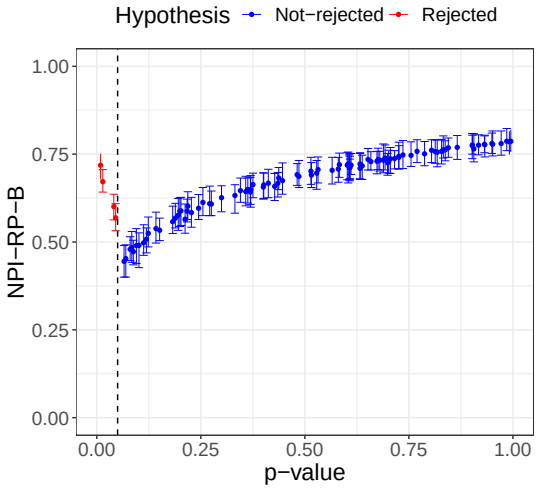
(b) ANOVA, $n = 6$



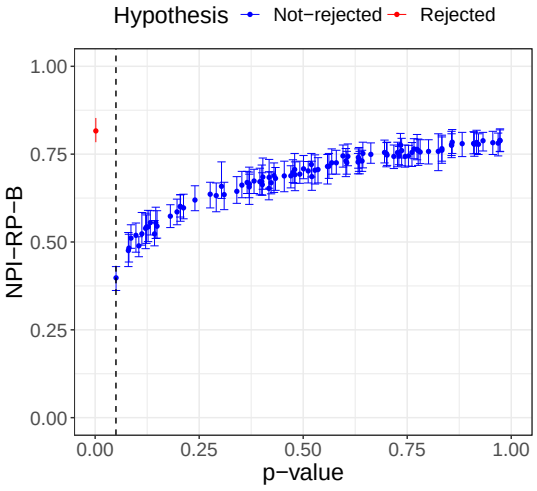
(c) KW, $n = 10$



(d) ANOVA, $n = 10$



(e) KW, $n = 20$



(f) ANOVA, $n = 20$

Figure 2.4: NPI-RP-B under H_0, with Case 2, $\alpha = 0.05$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
17.360	1.700×10^{-4}	R	0.599	0.939	0.961	0.961	0.975	18.978	7.130×10^{-6}	R	0.584	0.952	0.964	0.964	0.978
15.246	4.890×10^{-4}	R	0.526	0.945	0.959	0.959	0.970	14.606	5.019×10^{-5}	R	0.520	0.937	0.958	0.958	0.969
12.266	0.002	R	0.423	0.833	0.85	0.855	0.879	11.348	2.649×10^{-4}	R	0.457	0.871	0.892	0.892	0.915
10.692	0.005	R	0.369	0.761	0.788	0.790	0.817	8.052	0.002	R	0.374	0.779	0.806	0.807	0.838
9.092	0.011	R	0.314	0.716	0.750	0.750	0.780	6.620	0.005	R	0.329	0.728	0.758	0.758	0.791
8.519	0.014	R	0.294	0.622	0.654	0.655	0.681	5.223	0.012	R	0.279	0.626	0.655	0.653	0.691
6.947	0.031	R	0.240	0.558	0.593	0.594	0.623	4.321	0.024	R	0.242	0.551	0.582	0.581	0.617
5.907	0.052	NR	0.204	0.415	0.448	0.449	0.484	3.252	0.054	NR	0.194	0.432	0.467	0.469	0.501
4.338	0.114	NR	0.150	0.485	0.529	0.529	0.564	3.528	0.044	R	0.207	0.505	0.541	0.539	0.591
2.841	0.242	NR	0.098	0.598	0.624	0.623	0.661	2.044	0.149	NR	0.131	0.563	0.592	0.591	0.622

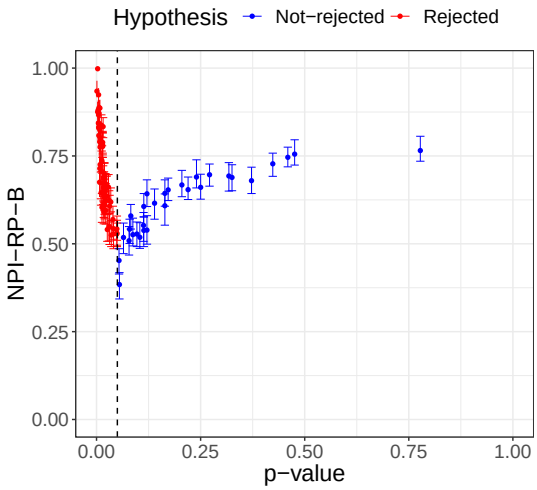
Table 2.13: RP under H_1, with Case 3, $n = 10$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 27) = 3.354$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
36.701	1.073×10^{-8}	R	0.622	0.999	1.000	1	1	42.850	4.379×10^{-12}	R	0.601	0.999	1.000	1	1
34.055	4.027×10^{-8}	R	0.577	0.998	1.000	1	1	43.432	3.474×10^{-12}	R	0.604	0.998	1.000	1	1
27.378	1.135×10^{-6}	R	0.464	0.989	0.996	0.996	1	21.535	1.081×10^{-7}	R	0.430	0.993	0.997	0.997	1
21.727	1.915×10^{-5}	R	0.368	0.940	0.959	0.959	0.975	16.922	1.702×10^{-5}	R	0.373	0.947	0.961	0.961	0.975
19.478	5.895×10^{-5}	R	0.330	0.933	0.948	0.949	0.964	14.280	9.366×10^{-5}	R	0.334	0.929	0.947	0.947	0.968
16.974	2.061×10^{-4}	R	0.288	0.917	0.934	0.934	0.955	14.841	6.479×10^{-6}	R	0.342	0.944	0.959	0.959	0.975
15.419	4.485×10^{-4}	R	0.261	0.887	0.905	0.906	0.924	10.958	9.400×10^{-5}	R	0.278	0.911	0.926	0.926	0.949
13.941	9.393×10^{-4}	R	0.236	0.870	0.896	0.896	0.917	8.499	5.886×10^{-4}	R	0.230	0.872	0.893	0.893	0.918
12.880	0.002	R	0.218	0.825	0.857	0.857	0.890	7.777	1.032×10^{-3}	R	0.214	0.791	0.819	0.818	0.852
10.771	0.005	R	0.183	0.778	0.805	0.805	0.841	7.565	1.220×10^{-3}	R	0.210	0.806	0.843	0.842	0.880

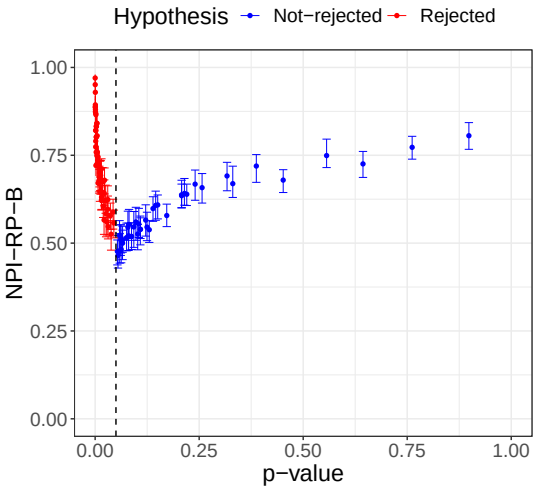
Table 2.14: RP under H_1, with Case 3, $n = 20$, $\chi^2_{2,0.05} = 5.99$, $F(0.05, 2, 57) = 3.159$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
10.034	0.040	R	0.346	0.656	0.689	0.690	0.724	2.583	0.062	NR	0.292	0.327	0.366	0.368	0.397
9.084	0.059	NR	0.313	0.333	0.365	0.365	0.403	2.800	0.048	R	0.309	0.599	0.646	0.647	0.677
8.856	0.065	NR	0.305	0.389	0.429	0.431	0.462	2.318	0.085	NR	0.271	0.374	0.412	0.414	0.454
7.088	0.131	NR	0.244	0.375	0.412	0.413	0.456	2.435	0.074	NR	0.280	0.330	0.363	0.362	0.395
6.748	0.150	NR	0.233	0.445	0.502	0.503	0.548	1.823	0.156	NR	0.226	0.443	0.497	0.499	0.548
5.647	0.227	NR	0.195	0.509	0.552	0.553	0.591	1.671	0.188	NR	0.211	0.465	0.510	0.511	0.548
4.912	0.297	NR	0.169	0.518	0.558	0.560	0.607	1.816	0.157	NR	0.225	0.453	0.494	0.495	0.529
3.161	0.531	NR	0.109	0.598	0.648	0.647	0.677	0.717	0.588	NR	0.103	0.590	0.633	0.633	0.678
1.166	0.884	NR	0.040	0.685	0.723	0.723	0.762	0.261	0.900	NR	0.040	0.663	0.698	0.699	0.735
0.142	0.998	NR	0.005	0.719	0.754	0.754	0.790	0.178	0.948	NR	0.028	0.684	0.709	0.708	0.736

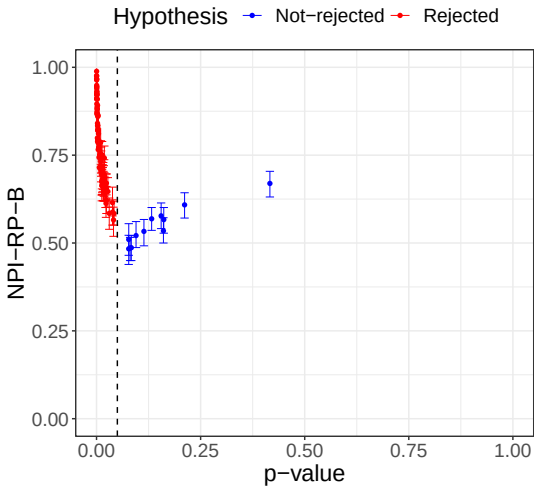
Table 2.15: RP under H_0, with Case 4, $n = 6$, $\chi^2_{4,0.05} = 9.49$, $F(0.05, 4, 25) = 2.759$



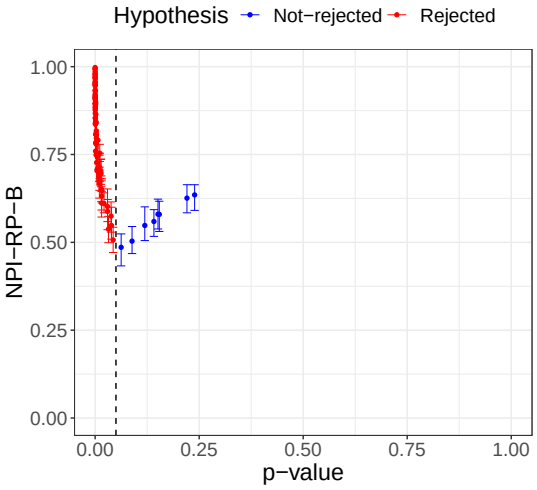
(a) KW, $n = 6$



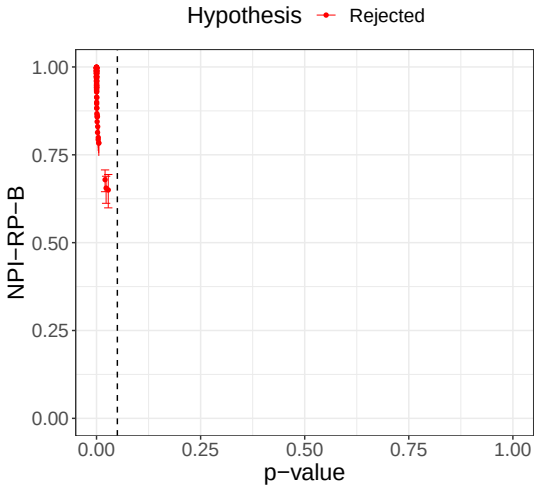
(b) ANOVA, $n = 6$



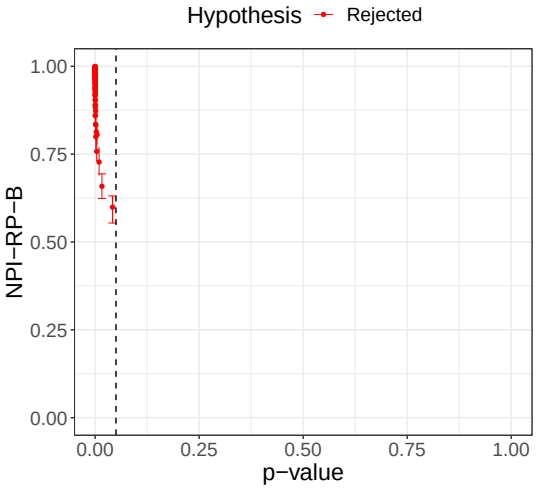
(c) KW, $n = 10$



(d) ANOVA, $n = 10$



(e) KW, $n = 20$



(f) ANOVA, $n = 20$

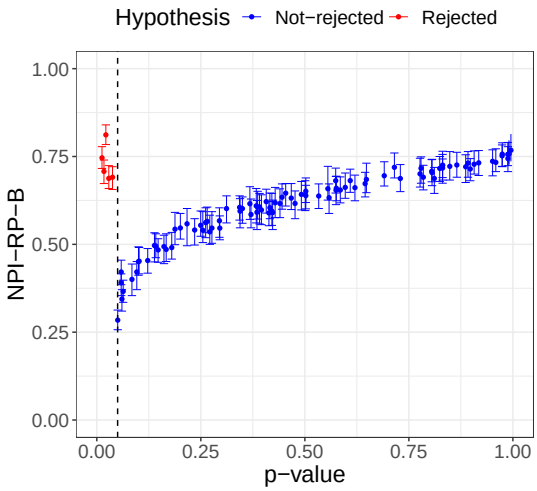
Figure 2.5: NPI-RP-B under H_1, with Case 3, $\alpha = 0.05$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
11.290	0.023	R	0.230	0.768	0.805	0.806	0.836	3.218	0.021	R	0.222	0.720	0.757	0.756	0.790
9.019	0.061	NR	0.184	0.290	0.330	0.330	0.362	2.208	0.083	NR	0.164	0.362	0.398	0.400	0.424
7.997	0.092	NR	0.163	0.367	0.402	0.401	0.440	1.870	0.132	NR	0.143	0.396	0.438	0.437	0.473
6.949	0.139	NR	0.142	0.411	0.440	0.440	0.476	1.742	0.157	NR	0.134	0.408	0.446	0.447	0.477
5.711	0.222	NR	0.117	0.448	0.478	0.476	0.509	1.783	0.149	NR	0.137	0.410	0.446	0.446	0.478
4.703	0.319	NR	0.096	0.457	0.503	0.503	0.542	1.742	0.158	NR	0.134	0.410	0.450	0.451	0.481
3.811	0.432	NR	0.078	0.509	0.556	0.557	0.601	0.587	0.674	NR	0.050	0.569	0.614	0.614	0.649
2.812	0.590	NR	0.057	0.573	0.603	0.604	0.637	0.788	0.539	NR	0.065	0.556	0.590	0.590	0.630
1.113	0.892	NR	0.023	0.633	0.676	0.676	0.714	0.313	0.867	NR	0.027	0.634	0.668	0.670	0.703
0.024	1.000	NR	4.802×10^{-4}	0.704	0.738	0.738	0.777	0.073	0.990	NR	0.006	0.685	0.715	0.715	0.751

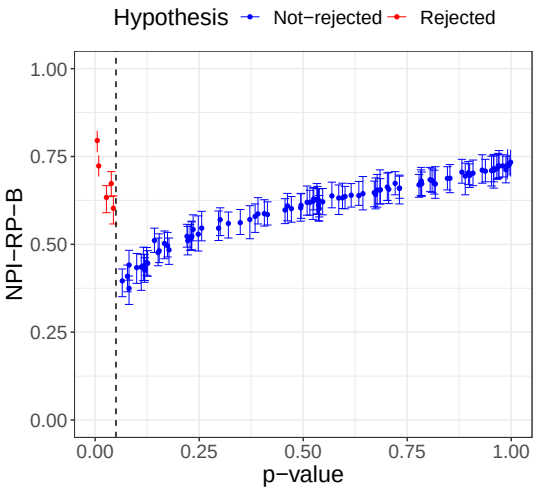
Table 2.16: RP under H_0, with Case 4, $n = 10$, $\chi^2_{4,0.05} = 9.49$, $F(0.05, 4, 45) = 2.579$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
13.298	0.010	R	0.134	0.792	0.827	0.828	0.856	3.816	0.006	R	0.138	0.793	0.822	0.822	0.852
9.527	0.049	R	0.096	0.677	0.712	0.712	0.739	2.477	0.049	R	0.094	0.643	0.681	0.681	0.708
9.410	0.052	NR	0.095	0.227	0.268	0.269	0.299	2.407	0.055	NR	0.092	0.281	0.317	0.318	0.349
8.094	0.088	NR	0.082	0.313	0.343	0.342	0.382	2.674	0.037	NR	0.101	0.663	0.701	0.701	0.734
6.743	0.150	NR	0.068	0.330	0.374	0.373	0.410	1.628	0.174	NR	0.064	0.359	0.396	0.395	0.430
5.043	0.283	NR	0.051	0.412	0.453	0.453	0.481	1.292	0.279	NR	0.052	0.429	0.472	0.472	0.511
3.808	0.433	NR	0.038	0.474	0.509	0.509	0.542	0.950	0.439	NR	0.038	0.479	0.531	0.532	0.564
2.851	0.583	NR	0.029	0.534	0.561	0.563	0.593	0.532	0.712	NR	0.022	0.567	0.601	0.601	0.636
1.522	0.823	NR	0.015	0.585	0.621	0.621	0.654	0.805	0.525	NR	0.033	0.528	0.558	0.556	0.598
0.089	0.999	NR	8.953×10^{-4}	0.661	0.701	0.701	0.749	0.106	0.980	NR	0.004	0.640	0.683	0.683	0.726

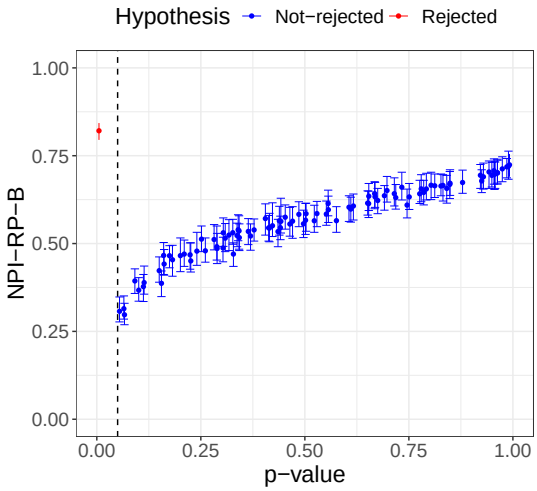
Table 2.17: RP under H_0, with Case 4, $n = 20$, $\chi^2_{4,0.05} = 9.49$, $F(0.05, 4, 95) = 2.467$



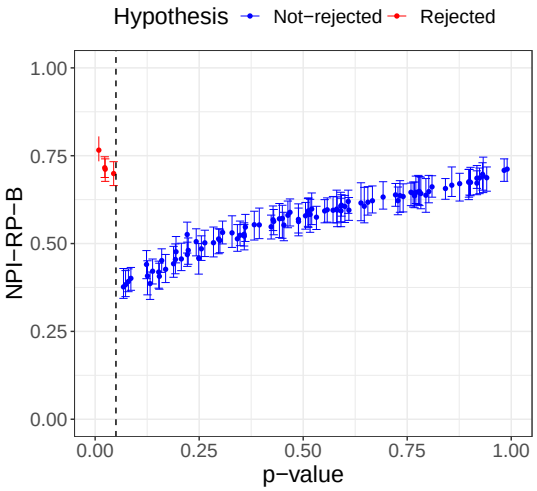
(a) KW, $n = 6$



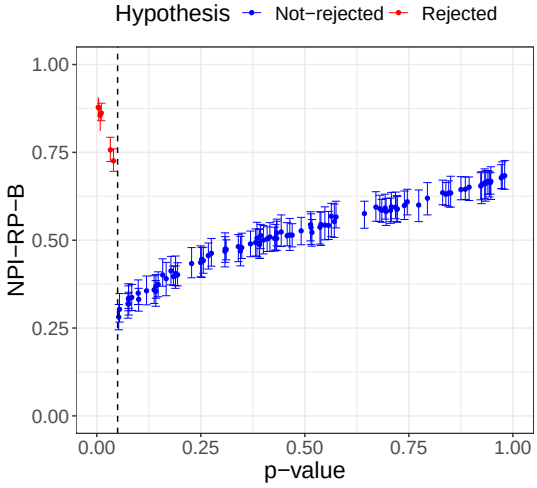
(b) ANOVA, $n = 6$



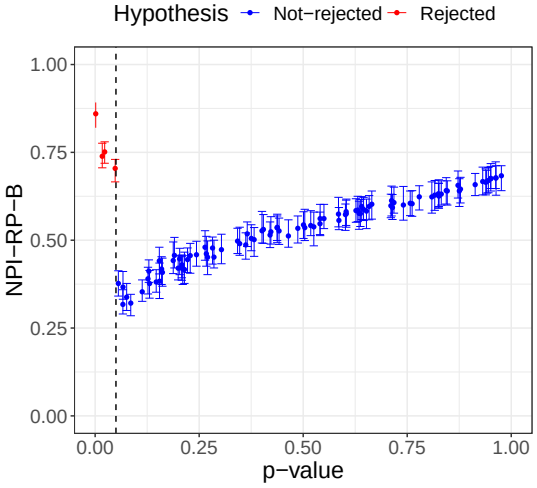
(c) KW, $n = 10$



(d) ANOVA, $n = 10$

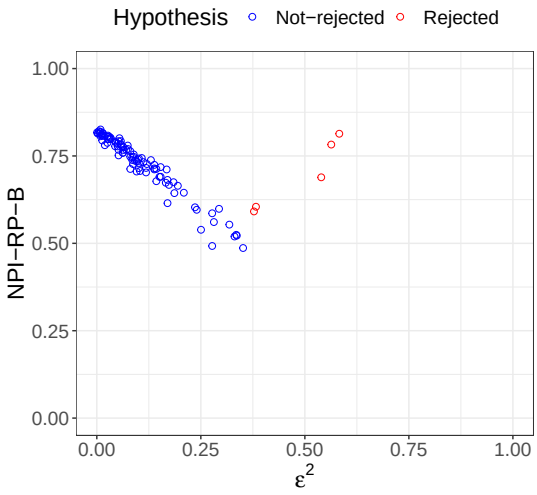


(e) KW, $n = 20$

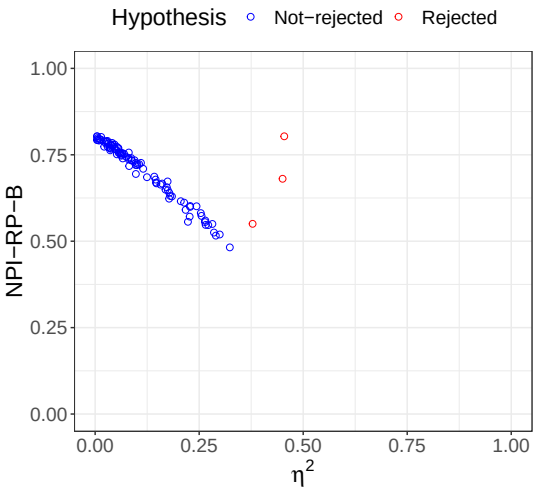


(f) ANOVA, $n = 20$

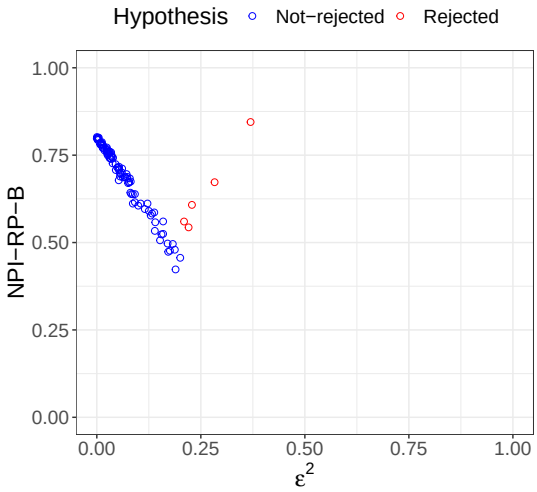
Figure 2.6: NPI-RP-B under H_0, with Case 4, $\alpha = 0.05$



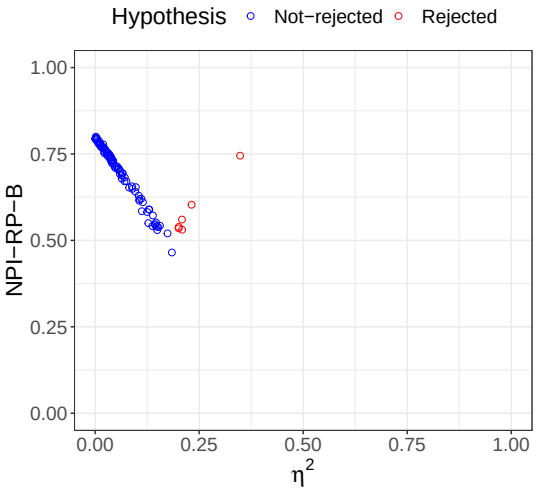
(a) KW, $n = 6$



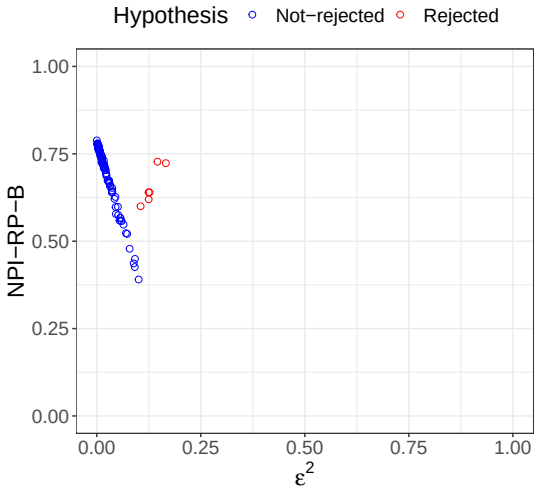
(b) ANOVA, $n = 6$



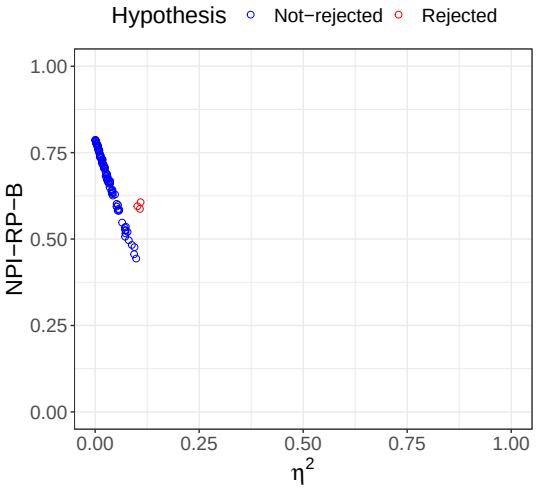
(c) KW, $n = 10$



(d) ANOVA, $n = 10$

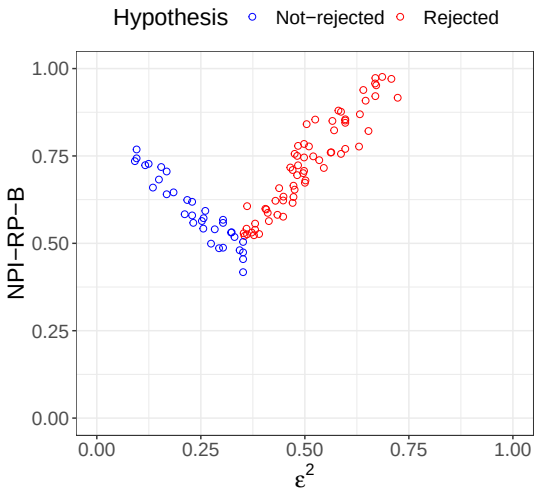


(e) KW, $n = 20$

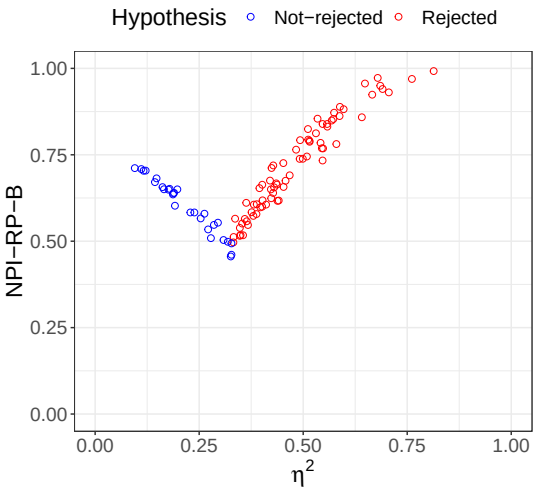


(f) ANOVA, $n = 20$

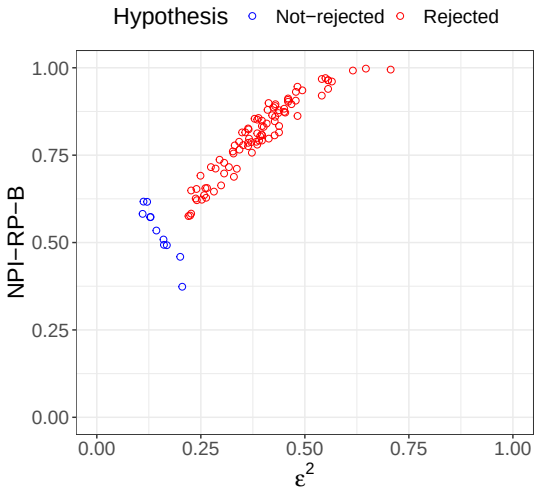
Figure 2.7: Simulations under H_0: NPI-RP-B vs effect size



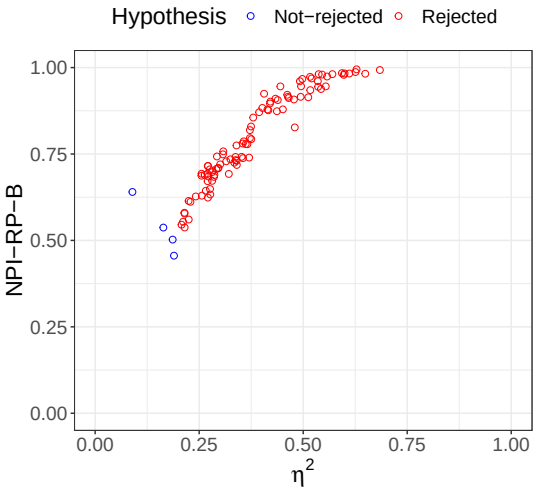
(a) KW, $n = 6$



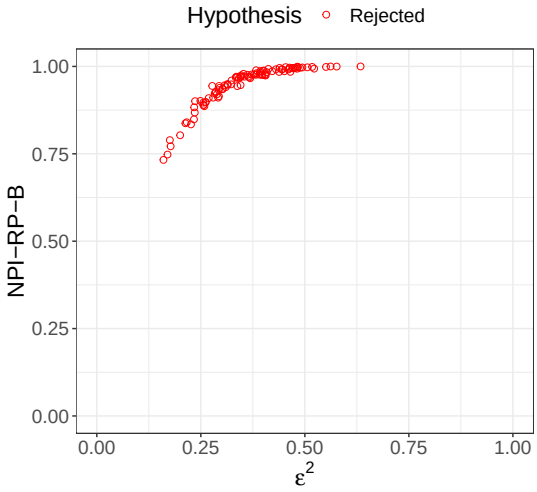
(b) ANOVA, $n = 6$



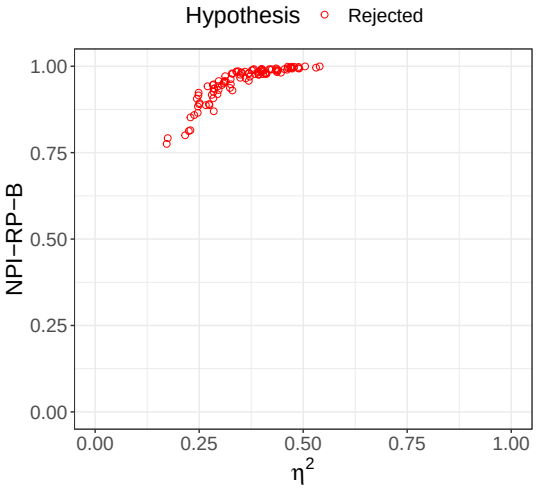
(c) KW, $n = 10$



(d) ANOVA, $n = 10$



(e) KW, $n = 20$



(f) ANOVA, $n = 20$

Figure 2.8: Simulations under H_1: NPI-RP-B vs effect size

2.6 Concluding remarks

This chapter explored the reproducibility probability for the KW test and the ANOVA test using NPI-RP-B. We provided a comparison of the NPI-RP for the KW test and the ANOVA test through simulation studies and data sets from the literature. The reproducibility probability of the tests is explored with samples of the same size and significance level as in the actual test, because this approach seems logical and natural from the perspective of theoretical reproducibility within a frequentist statistical framework.

In this chapter, the NPI-RP estimates for both the KW test and the ANOVA test are quite similar. The estimates of the NPI reproducibility probability for both tests tend to increase as the p-value moves away from the test threshold, regardless of the decision on H_0. The results presented for the estimates of the NPI reproducibility probability show consistency with previous NPI studies of test reproducibility [2, 32, 33, 75, 98]. In terms of estimating reproducibility probability, there is a straightforward argument that if the distribution under the null hypothesis of the test statistic is (about) symmetric, then a worst case scenario would provide reproducibility probability of (about) 0.5 [51, 94].

The use of NPI-RP-B approach to estimate reproducibility probability avoids the hard calculations of the lower and upper reproducibility probabilities, and it is a suitable approach for large sample sizes. In this chapter, the NPI-RP-B is considered to approximate the RP, as deriving explicit analytical formulas for the exact lower and upper ANOVA test is not trivial. Further exploration of the exact approach of NPI-RP for the KW test and the one-way ANOVA test is interesting topic for future research. NPI-RP for the KW test and the one-way ANOVA test can be explored by applying the Paramtric Predictive Bootsrap method introduced in [2]. Another idea for future research is to explore the NPI reproducibility of the two-way ANOVA test, and its nonparametric analogue the Friedman test [108].

Chapter 3

Reproducibility of Ordered Alternatives Tests

3.1 Introduction

In statistical hypothesis testing, the alternative hypothesis can be either directional or nondirectional, depending on the research question that needs to be addressed and the context of the research. Chapter 2 introduced the NPI reproducibility probability for two general alternatives tests, namely, the KW test and the ANOVA test. The general alternatives tests are used for nondirectional alternatives, specifically to determine whether there are statistically significant differences between the means for three or more independent groups. In some applications it is of interest to test for specific patterns or trends of the differences between the means. The ordered alternatives is sometimes more meaningful than the nondirectional alternatives, and it can help researchers make more informed conclusions about the differences in the groups means.

This chapter contributes to statistical test reproducibility by considering NPI reproducibility probability for ordered alternatives, namly, for the Jonckheere-Terpstra (JT) test. The Jonckheere-Terpstra (JT) test is a nonparametric test that can be used to test an ordered alternative hypothesis for three or more independent groups. NPI approach for reproducibility probability involves deriving the exact lower and upper reproducibility probabilities, as in Section 1.5. However, deriving the exact lower and upper RP is not trivial for the JT test. BinHimd [18] and Simkus [97] encountered the same challenge while exploring the NPI reproducibility probability for the Kolmogorov Smirnov test and the t-test. They addressed this issue by applying the NPI bootstrap method, introduced in Section 1.5.3, to compute an approximate NPI reproducibility probability. This chapter is mainly focused on calculating estimates for NPI reproducibility probabilities using NPI bootstrap which provides a point estimate for the NPI

reproducibility probability rather than lower and upper reproducibility probabilities.

Section 3.2 provides a brief review of the ordered alternatives tests. The Jonckheere-Terpstra test for ordered alternatives is presented in Section 3.3. In Section 3.4, the NPI-B method is introduced to investigate the reproducibility probability for the Jonckheere-Terpstra test. In Section 3.5, simulation studies were carried out to get an insight into how different distributions and sample sizes impact the reproducibility probability for the JT test. Section 3.6 presents some concluding remarks for this chapter.

3.2 Ordered alternatives tests

A common problem in statistical analysis is to decide whether several groups should be regarded as coming from the same population. The most general form of the alternative is

$$H_1 : \text{at least one } \mu_i \text{ is different.} \tag{3.1}$$

where μ_i is the location parameter of the ith group. However, in some applications, it possible to be more precise in the specification of the alternatives. Therefore, instead of the unrestricted alternative mentioned above, the ordered alternative can be considered. The term ordered alternative refers to a monotonic trend, either increasing or decreasing, in the alternative hypothesis.

Jonckheere [61] provided an example of an application, to analyze an experiment performed to determine the effect of different degrees of stress on the performance of some task of manual dexterity, where the data were obtained from groups of subjects working under high, medium, low and minimal stress. The null hypothesis being that stress has no effect on performance, and the alternative hypothesis is that increasing stress produces an increasing effect. Other applications of this nature occur in social sciences experiments where participants can be grouped according to social class, degree of satisfaction, etc. Such variables can be ranked based on their expected effect.

Several nonparametric tests are available in the literature to test the equality of locations against the ordered alternatives. The Jonckheere-Terpstra test is the most common test for ordered alternatives when the data conform to the format for one-way analysis of variance, and was proposed by Jonckheere [61] and Terpstra [101]. Tryon and Hettmansperger [106] proposed the modified Jonckeere-Terpstra test, where the test statistic is computed based on the Mann-Whitney statistic [73]. Cuzick [36] developed an extension of the Wilcoxon rank-sum test to handle the situation in which test for trend for three or more groups is desired. Page [86] proposed another ranked based nonparametric test for ordered alternatives that is appropriate in two-way analysis of variance situations.

Some attention has also been given to explicitly ordered alternative hypothesis in parametric analysis. Bartholomew [11, 12] explored the problem of assigning weights to classes and using likelihood ratio tests. Other related work and approaches constructing test statistics for ordered alternatives were also presented by [5, 84, 90], but they will not be addressed in this chapter. This chapter will focus on investigating the reproducibility of the Jonckheere-Terpstra test.

3.3 Jonckheere-Terpstra (JT) test

The Jonckheere-Terpstra test is a nonparametric test which was proposed independently by Jonckheere [61] and Terpstra [101] to test an ordered alternative hypothesis. The alternative hypothesis asserts that there is a specific increasing (or decreasing) order of three or more population location parameters with at least one mean is different [59, 99]. The null and alternative hypotheses can be expressed in terms of the k population means as follows

$$H_0 : \mu_1 = \mu_2 = ... = \mu_k \tag{3.2}$$

$$H_1 : \mu_1 \leq \mu_2 \leq ... \leq \mu_k \tag{3.3}$$

The samples must be labeled prior to data collection in such a way that the experimenter expects any deviation from H_0 to be in the particular direction associated with H_1. We emphasize, however, that the labeling of the samples must correspond completely to samples implicit in the nature of the experimental design and not the observed sample observations [59]. To compute the Jonckheere–Terpstra test statistic, J, we need to calculate the $k(k\text{-}1)/2$ Mann–Whitney counts U_{uv}. So, U_{uv} is the number of observations from sample u which are smaller than the observations from sample v [59]. Formally, the U_{uv} is given by

$$U_{uv} = \sum_{i=1}^{n_u} \sum_{j=1}^{n_v} 1(X_{iu} < X_{jv}), \text{ for all } 1 \leq u < v \leq k, \tag{3.4}$$

where the values of sample u are denoted by X_{iu} and the values of sample v are denoted by X_{jv}, and $1(X_{iu} < X_{jv})$ is an indicator function that equals 1 if $X_{iu} < X_{jv}$ and 0 otherwise. The test statistic, J, is the sum of these $k(k\text{-}1)/2$ Mann–Whitney counts,

$$J = \sum_{u=1}^{v-1} \sum_{v=2}^{k} U_{uv} \tag{3.5}$$

Then the null hypothesis is rejected against the alternative, at level of significance α, if $J \geq J_\alpha$, where J_α is the $\alpha-$upper quantile of the null distribution of J [59].

For large sample sizes, and under the null hypothesis, the statistic J is asymptotically

Normally distributed with the following mean and variance:

$$E(J) = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4} \tag{3.6}$$

$$\sigma^2(J) = \frac{N^2(2N+3) - \sum_{i=1}^{k} n_i^2(2n_i+3)}{72} \tag{3.7}$$

where N is the total number of observations and n_i is the number of data observations in group i [59]. The standardized version of J is given by

$$J^* = \frac{J - E(J)}{\sigma(J)} \tag{3.8}$$

The null hypothesis is then rejected, at level of significant α, if $J^* \geq Z_\alpha$. The power of the test is approximated by:

$$Power = 1 - \Phi([\mu_{(0)} - \mu_{(A)}]/\sigma_{(0)} + Z_\alpha) \tag{3.9}$$

where Z_α is the α-upper quantile of the standard Normal distribution. Under both the null and alternative hypothesis, $\mu_{(0)}$ and $\mu_{(A)}$ are the expectations of J, respectively, as follows,

$$\mu_{(0)} = \sum_{i<j} \frac{1}{2} n_i n_j; \text{ for all } i \neq j \tag{3.10}$$

$$\mu_{(A)} = \sum_{i<j} \Phi(\delta/\sqrt{2}) n_i n_j \tag{3.11}$$

with δ represents the amount by which a location parameter for the population from which the sample exceeds that of the other sample [65]. The null variance of J is

$$\sigma^2_{(0)} = \frac{1}{27}[N(N+1)(2N+1) - \sum_{i=1}^{k} n_i(n_i+1)(2n_i+1)] \tag{3.12}$$

The power can be computed using the function `terpstrapower` from the R package `MultNonParam` [65].

3.4 NPI-RP-B for the JT test

This section investigates the reproducibility probability for the Jonckheere-Terpstra (JT) test, which was briefly reviewed in Section 3.3. As mentioned earlier in this chapter, deriving exact lower and upper reproducibility probability for the JT test is not trivial. This can be resolved by using the NPI-RP-B method which was introduced in Section 1.5.3. BinHimd [18] proposed the use of the NPI-B method as a heuristic method to approximate the NPI reproducibility probability, as it avoids the complex calculations required by the exact NPI reproducibility

probability approach. The explicitly predictive nature of NPI-B provides a natural formulation of inferences on reproducibility of statistical tests. It is important to emphasize that we focus on the conclusion of the future test with regard to the null hypothesis, given the actual data of the original test. It is worth noting that the NPI framework for statistical tests reproducibility does not require that the sample sizes in the initial and the future tests to be equal. However, it is a natural assumption to make for the sake of reflecting reproducibility. In this thesis, we will restrict our attention to the case where the number of future observations are equal to the number of the original data observations.

Algorithm 2 introduced in Section 1.5.3 is applied. The inputs into Algorithm 2 are the k original samples, their corresponding sample sizes, the number of runs T and the number of bootstrapped samples per run B. Summary statistics including the minimum, mean, median and maximum, of $RP_1, RP_2, ..., RP_T$ were calculated. In this chapter, Algorithm 2, will be implemented with both finite and infinite intervals, using Approach I and II, introduced in Section 2.3. In this thesis, the mean of $RP_1, RP_2, ..., RP_T$ is the reproducibility probability estimate, and is referred to as NPI-RP-B value.

Throughout this thesis, the results in the tables were rounded to three decimal digits and precise value 1 is presented without additional decimals, so the values 1.000 are less than 1 but rounded up. Furthermore, the test outcome is either to reject (R) or to not reject (NR) the null hypothesis. Section 3.5 presents the results of simulation studies for different scenarios, under H_0 and under H_1, with varying sample sizes and number of groups. Further, in this chapter, we consider an increasing ordered alternative with $H_1 : \mu_1 \leq \mu_2 \leq ... \leq \mu_k$. We do not illustrate cases with $H_1 : \mu_1 \geq \mu_2 \geq \ldots \geq \mu_k$, as these follow by symmetry and therefore have similar behavior.

A common observed pattern from the previous NPI-RP studies of test reproducibility is expected: that the original test statistic close to the threshold between rejection and non-rejection of the null hypothesis is linked to low reproducibility probability. This pattern is also observed in Chapter 2. In practice researchers often focus on the rejection of the null hypothesis when studying the reproducibility of statistical tests, as it is a major concern, particularly when significant results cannot be replicated in subsequent studies. This tends to be the most important scenario in medical research, particularly in relation to the introduction of new medications. Nevertheless, for a comprehensive understanding, we believe that the reproducibility of statistical tests that did not yield significant results is also important. Therefore, we consider reproducibility probability for both cases of rejection and non-rejection of the null-hypothesis.

Example 3.1. This example is introduced to study the NPI reproducibility probability for

X	Y	Z	J	p-value	H_0	Min	Mean	Median	Max
1,2,3	4,5,6	7,8,9	27	0.001	R	0.982	0.991	0.991	0.998
1,2,3	4,5,8	6,7,9	25	0.005	R	0.692	0.731	0.731	0.771
1,2,3	4,6,8	5,7,9	24	0.011	R	0.627	0.667	0.667	0.705
1,2,6	3,4,7	5,8,9	23	0.021	R	0.363	0.401	0.401	0.445
1,3,7	2,4,5	6,8,9	22	0.037	R	0.367	0.400	0.401	0.438
1,3,6	2,4,8	5,7,9	21	0.061	NR	0.710	0.739	0.740	0.765
1,2,4	5,6,9	3,7,8	21	0.061	NR	0.637	0.675	0.676	0.709
1,2,5	4,7,8	3,6,9	20	0.095	NR	0.686	0.719	0.721	0.748
1,3,7	2,4,8	5,6,9	20	0.095	NR	0.760	0.786	0.785	0.817
1,5,7	2,3,6	4,8,9	19	0.139	NR	0.756	0.787	0.787	0.820
1,2,5	6,8,9	3,4,7	17	0.260	NR	0.877	0.898	0.898	0.918
2,4,6	1,8,9	3,5,7	15	0.416	NR	0.957	0.972	0.972	0.985
1,6,9	4,5,8	2,3,7	10	0.806	NR	0.930	0.947	0.947	0.961
5,8,9	2,6,7	1,3,4	4	0.989	NR	0.997	1.000	1	1
7,8,9	5,6,7	1,2,3	0	0.999	NR	1	1	1	1

Table 3.1: RP for the JT test with $H_1 : \mu_x \leq \mu_y \leq \mu_z$, $n = 3$, $\alpha = 0.05$, $J_{0.0369} = 22$

the JT test. We consider artificial data sets of ranks for $k = 3$ groups of sizes $n = 3, 4, 5$, to illustrate how the original samples ranks impact the NPI-RP values. To test the hypothesis $H_0 : \mu_x = \mu_y = \mu_z$ against an increasing ordered alternative hypothesis H_1: $\mu_x \leq \mu_y \leq \mu_z$, the level of significance is set at $\alpha = 0.05$. Notice that for the case with $n = 3$ in Table 3.1, with $\alpha = 0.05$, due to the discrete nature of the test statistics the nominal level is 0.0369 which leads to the critical value $J_{0.0369} = 22$. Thus, the null hypothesis is rejected if $J \geq 22$. For the case with $n = 4$ in Table 3.2, the nominal level is 0.0463 and the test decision rule is to reject the null hypothesis if $J \geq 36$. For the case with $n = 5$ in Table 3.3, the nominal level is 0.0456, which leads to the null hypothesis being rejected if $J \geq 54$.

The NPI-RP-B approach introduced in Section 1.5.3 is considered for the ranks given in Tables 3.1, 3.2 and 3.3. Algorithm 1 has been applied using Approach I, introduced in Section 2.3, with $B = 1000$ and $T = 100$. In Approach I, the lower limit is taken to be smallest value of the group minus the maximal distance between consecutive points, and the upper limit is taken to be equal to largest value of the group plus the maximal distance between consecutive points. For the estimates of RP for the ranks in the first row in Tables 3.1, 3.2 and 3.3, theoretically when the data are perfectly ordered the upper reproducibility probability will be equal to 1,

X	Y	Z	J	p-value	H_0	Min	Mean	Median	Max
1,2,3,4	5,6,7,8	9,10,11,12	48	2.886×10^{-5}	R	0.999	1.000	1	1
1,2,3,5	4,6,7,10	8,9,11,12	45	0.001	R	0.919	0.941	0.940	0.958
1,2,3,8	4,6,9,10	5,7,11,12	39	0.015	R	0.450	0.483	0.483	0.530
1,2,3,9	4,6,8,10	5,7,11,12	38	0.023	R	0.440	0.470	0.470	0.515
1,2,3,12	4,5,7,9	6,8,10,11	37	0.033	R	0.390	0.428	0.427	0.472
1,2,3,12	4,6,7,9	5,8,10,11	36	0.046	R	0.358	0.398	0.397	0.441
1,2,5,8	4,6,9,10	3,7,11,12	36	0.046	R	0.365	0.402	0.402	0.442
1,2,5,9	4,6,7,11	3,8,10,12	35	0.063	NR	0.619	0.657	0.658	0.691
1,3,5,8	4,6,7,12	2,9,10,11	34	0.084	NR	0.631	0.669	0.671	0.708
1,3,5,9	4,6,7,11	2,8,10,12	34	0.084	NR	0.642	0.688	0.691	0.717
1,3,5,10	4,6,7,12	2,8,9,11	32	0.140	NR	0.700	0.754	0.754	0.788
1,3,9,10	4,6,7,12	2,5,8,11	26	0.416	NR	0.825	0.866	0.866	0.890
1,9,10,11	4,6,7,12	2,3,5,8	15	0.916	NR	0.963	0.976	0.977	0.987
9,10,11,12	5,6,7,8	1,2,3,4	0	1	NR	1	1	1	1

Table 3.2: RP for the JT test with $H_1 : \mu_x \leq \mu_y \leq \mu_z$, $n = 4$, $\alpha = 0.05$, $J_{0.0463} = 36$

but since an approximation method is applied, the estimates of NPI reproducibility probability are close to 1 [18]. In most cases where different ranks per sample yield the same value of the test statistic J, the estimates of RP differ, that is because they depend on the actual ranks per sample and not just on the value of the test statistic, as for the ranks with $J = 20$ and $J = 36$ in Tables 3.1 and 3.2, respectively. It is clear that, as expected, the reproducibility probability is small when J is close to the threshold, substantially smaller than 0.5. In such cases, reproducibility tends to be lower in the case of rejection than for non-rejection. Further, the reproducibility tends to be larger the further away the original test statistic J is from the threshold.

X	Y	Z	J	p-value	H_0	Min	Mean	Median	Max
1,2,3,4,5	6,7,8,9,10	11,12,13,14,15	75	1.321×10^{-6}	R	0.999	1.000	1	1
1,2,3,6,7	4,5,8,9,11	10,12,13,14,15	70	9.779×10^{-5}	R	0.969	0.981	0.982	0.991
1,2,3,6,7	4,5,8,11,12	9,10,13,14,15	67	0.001	R	0.914	0.932	0.933	0.952
1,2,6,7,8	3,4,5,11,12	9,10,13,14,15	62	0.004	R	0.732	0.765	0.766	0.793
1,2,6,7,8	3,4,11,12,13	5,9,10,14,15	57	0.021	R	0.456	0.497	0.499	0.538
1,2,6,7,12	3,4,8,11,13	5,9,10,14,15	55	0.036	R	0.368	0.409	0.411	0.439
1,2,6,7,13	3,4,8,11,12	5,9,10,14,15	54	0.046	R	0.356	0.394	0.395	0.423
1,2,6,7,14	3,4,8,11,12	5,9,10,13,15	53	0.057	NR	0.610	0.636	0.635	0.674
1,2,6,7,15	3,4,8,11,12	5,9,10,14,13	52	0.071	NR	0.629	0.657	0.655	0.697
1,2,6,9,15	3,4,8,11,12	5,7,10,14,13	49	0.126	NR	0.672	0.710	0.709	0.751
1,3,11,14,15	2,4,8,12,13	5,6,7,9,10	33	0.698	NR	0.920	0.935	0.935	0.959
1,11,12,14,15	2,4,8,10,13	3,5,6,7,9	22	0.954	NR	0.971	0.982	0.983	0.990
7,10,11,12,13	4,5,8,14,15	1,2,3,6,9	17	0.988	NR	0.998	1.000	1	1
11,12,13,14,15	4,7,8,9.10	1,2,3,5,6	2	1	NR	1	1	1	1
11,12,13,14,15	6,7,8,9,10	1,2,3,4,5	0	1	NR	1	1	1	1

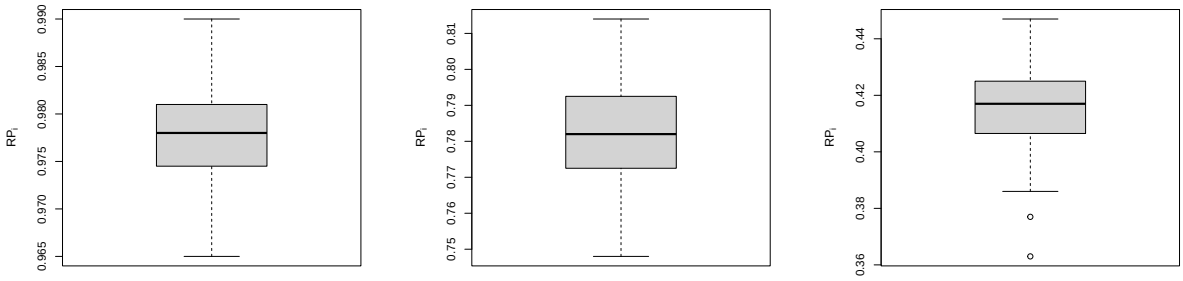
Table 3.3: RP for the JT test with $H_1 : \mu_x \leq \mu_y \leq \mu_z$, $n = 5$, $\alpha = 0.05$, $J_{0.0456} = 54$

Case	k	**Simulation**
1	3	$N(0, 1)$
2	3	$X \sim N(0, 1)$, $Y \sim N(1, 1)$, $Z \sim N(2, 1)$
3	3	$X \sim N(0.1, 1)$, $Y \sim N(0.1, 1)$, $Z \sim N(0.2, 1)$
4	3	$X \sim N(0.1, 1)$, $Y \sim N(0.2, 1)$, $Z \sim N(0.3, 1)$
5	3	Gamma(2, 1)
6	5	$N(0, 1)$
7	5	$X \sim N(0, 1)$, $Y \sim N(0.1, 1)$, $Z \sim N(0.2, 1)$, $V \sim N(0.3, 1)$, $W \sim N(0.4, 1)$

Table 3.4: Simulation cases for the JT test

3.5 Simulation study

This section studies the reproducibility probability for the JT test via simulations, where reproducibility is calculated using Algorithm 2. The NPI-RP-B method is preformed using Approach II. Data were simulated under H_0 and under H_1, from symmetric and skewed distributions, as listed in Table 3.4. To study the impact of the number of groups and the sample size on the reproducibility probability, the simulation is considered with the number of groups $k = 3, 5$ and the sample size $n = 6, 20$. The null hypothesis is $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ and the alternative



(a) $J = 34$ (b) $J = 63$ (c) $J = 75$

Figure 3.1: RP_i for $i = 1, ..., T$ for selected J from Table 3.5, $T = 100$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
77	0.034	R	0.378	0.414	0.415	0.454	58	0.391	NR	0.814	0.839	0.839	0.867
75	0.049	R	0.300	0.335	0.335	0.373	50	0.640	NR	0.915	0.941	0.942	0.965
74	0.058	NR	0.574	0.611	0.611	0.651	41	0.860	NR	0.968	0.981	0.981	0.989
74	0.058	NR	0.617	0.653	0.651	0.698	39	0.893	NR	0.961	0.974	0.974	0.985
73	0.068	NR	0.592	0.631	0.632	0.667	34	0.951	NR	0.965	0.978	0.978	0.990
72	0.080	NR	0.591	0.627	0.628	0.657	32	0.966	NR	0.977	0.986	0.986	0.996
71	0.093	NR	0.593	0.624	0.623	0.664	29	0.981	NR	0.986	0.994	0.994	1
68	0.140	NR	0.752	0.784	0.784	0.827	24	0.994	NR	0.996	0.999	0.999	1
63	0.250	NR	0.748	0.783	0.782	0.814	20	0.998	NR	0.997	0.999	1	1
61	0.303	NR	0.821	0.848	0.848	0.876	13	1.000	NR	0.996	0.999	0.999	1

Table 3.5: RP for the JT test under H_0, with Case 1, $n = 6$, $J_{0.0490} = 75$

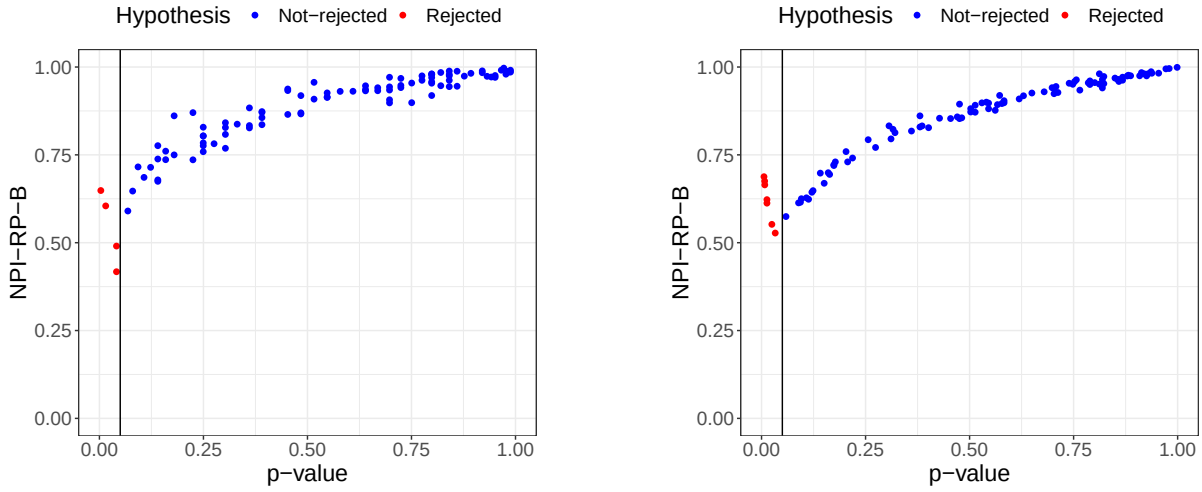
hypothesis is H_1: $\mu_1 \leq \mu_2 \leq ... \leq \mu_k$. The level of significance is set at $\alpha = 0.05$.

The inputs for the simulation study are as follows: Algorithm 2 is applied with $B = 1000$ and $T = 100$. For each run, one sample for each group k is generated from each of the cases given in Table 3.4, the JT test is performed on the these samples, and the test outcomes are obtained and the RP estimates for the JT test are calculated using Algorithm 1. These results can be illustrated by the case in Table 3.5, where J=63 in the original test. The JT test is performed on the three original samples which are drawn from $N(0, 1)$, and the resulting test statistic is J=63. In this instance, the null hypothesis is not rejected when compared to the threshold value 75. The second step is to draw an NPI-B sample from each of the three original samples and apply the JT test on these NPI-B samples to get the value of the test statistic J, and this step is repeated with $B = 1000$. Then, the estimate of NPI reproducibility probability is computed which is equal to the proportion of times H_0 is not rejected in the 1000 NPI-B

samples. Finally, we repeat the second step $T = 100$ times and obtain $RP_1, RP_2, ..., RP_{100}$, shown in Figure 3.1 where each RP value is the proportion of times H_0 is not rejected in 1000 NPI-B samples. The minimum, mean, median and maximum of these RP_i for $i = 1, 2, ..., 100$, are 0.748, 0.783, 0.782 and 0.814, respectively. The RP_i estimates for data sets with $J = 34$ and $J = 75$ presented in Table 3.5 are also visualized in Figure 3.1. In Tables 3.5 - 3.18, the reproducibility probability estimates have been reported for 20 simulated data sets for each scenario.

The relationship between NPI-RP-B and the p-value for the JT test is examined in the simulations. We use the p-value for better visualization of figures rather than the critical value because each simulation scenario has a different critical value, given the variations in the sample sizes and the number of groups considered. Although the p-values and critical values are two different approaches, they ultimately yield the same conclusion regarding whether the null hypothesis is rejected or not. Note that the level of significance $\alpha = 0.05$ is represented on the figures by a vertical line. For simulations under H_1 in Figures 3.3, 3.4, 3.5 and 3.8, with $n = 6$, the power of the JT test for each scenario is 0.864, 0.069, 0.094 and 0.182, respectively. Increasing the size of samples to $n = 20$ leads to increasing the power of the test to 1.000, 0.090, 0.151 and 0.390, respectively. This leads to more cases where the null hypothesis is rejected, as for higher power the test is more likely to correctly reject the null hypothesis when the alternative hypothesis is true. It is clear that, as expected, the reproducibility probability is small when the observed p-value is close to the threshold 0.05. There is a tendency for RP estimates to be lower in cases of rejection than in non-rejection, substantially smaller than 0.5. The reason for that is the presence of direction in the alternatives. The RP estimates tend to increase with increasing distance between the observed p-value and the threshold $\alpha = 0.05$, regardless of the decision about H_0. Moreover, RP for non-rejection cases for larger n becomes relatively lower compared to non-rejection cases for small n. Conversely, for the cases of rejection the reproducibility with larger n becomes relatively higher than for small n. Similar results have been observed in previous NPI-RP studies [2, 97].

It is noticed that the variability of NPI-RP-B for the JT test is reduced with increasing the size of samples. The median and the mean of RP_i are very close, that is some indication of reasonably symmetric distributions of the RP_i values for each simulated data set, where $i = 1, ..., 100$. Further simulations were performed for data generated under H_0 and H_1 for $k = 3$ and unequal sample sizes, the results are presented in the Appendix B.



(a) $n = 6$ (b) $n = 20$

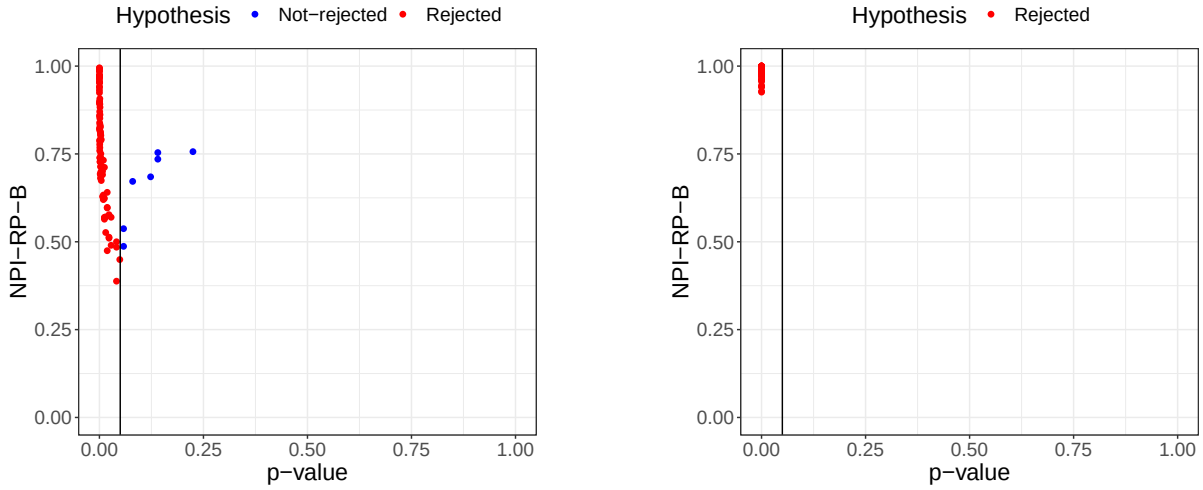
Figure 3.2: NPI-RP-B for the JT test under H_0, with Case 1, $\alpha = 0.05$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
729	0.041	R	0.459	0.499	0.500	0.534	594	0.535	NR	0.874	0.896	0.897	0.916
722	0.050	R	0.410	0.467	0.467	0.507	586	0.578	NR	0.895	0.914	0.914	0.936
713	0.064	NR	0.522	0.568	0.569	0.602	552	0.744	NR	0.914	0.936	0.936	0.959
698	0.093	NR	0.579	0.617	0.617	0.669	552	0.744	NR	0.941	0.957	0.956	0.977
692	0.108	NR	0.635	0.663	0.662	0.700	530	0.830	NR	0.941	0.959	0.959	0.974
666	0.188	NR	0.667	0.711	0.711	0.750	508	0.895	NR	0.967	0.978	0.977	0.993
647	0.265	NR	0.744	0.781	0.781	0.813	489	0.935	NR	0.978	0.986	0.987	0.994
627	0.360	NR	0.816	0.841	0.841	0.869	456	0.975	NR	0.977	0.989	0.990	0.995
600	0.503	NR	0.858	0.882	0.882	0.904	422	0.993	NR	0.997	0.999	1	1
600	0.503	NR	0.866	0.891	0.891	0.913	416	0.994	NR	0.993	0.998	0.998	1

Table 3.6: RP for the JT test under H_0, with Case 1, $n = 20$, $J_{0.0497} = 722$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
105	1.049×10^{-6}	R	0.982	0.990	0.991	0.996	79	0.023	R	0.501	0.545	0.545	0.589
102	8.103×10^{-6}	R	0.986	0.993	0.993	0.997	78	0.028	R	0.475	0.516	0.517	0.562
94	3.486×10^{-4}	R	0.932	0.951	0.951	0.966	77	0.034	R	0.406	0.453	0.455	0.488
94	3.486×10^{-4}	R	0.887	0.919	0.920	0.939	77	0.034	R	0.386	0.427	0.429	0.461
90	0.001	R	0.930	0.951	0.952	0.966	76	0.041	R	0.417	0.447	0.449	0.483
89	0.002	R	0.825	0.867	0.868	0.892	75	0.049	R	0.367	0.418	0.417	0.447
87	0.003	R	0.679	0.714	0.714	0.749	74	0.058	NR	0.594	0.624	0.623	0.672
85	0.006	R	0.615	0.648	0.651	0.679	72	0.080	NR	0.636	0.663	0.662	0.702
83	0.010	R	0.586	0.620	0.620	0.653	68	0.140	NR	0.677	0.709	0.709	0.748
80	0.019	R	0.564	0.594	0.595	0.628	56	0.453	NR	0.885	0.904	0.904	0.925

Table 3.7: RP for the JT test under H_1, with Case 2, $n = 6$, $J_{0.0490} = 75$



(a) $n = 6$

(b) $n = 20$

Figure 3.3: NPI-RP-B for the JT test under H_1, with Case 2, $\alpha = 0.05$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
1097	2.187×10^{-14}	R	0.999	1.000	1	1	983	2.157×10^{-8}	R	0.995	0.999	0.999	1
1085	1.388×10^{-13}	R	0.999	1.000	1	1	983	2.157×10^{-8}	R	0.994	0.997	0.997	1
1062	3.482×10^{-12}	R	0.999	1.000	1	1	963	1.275×10^{-7}	R	0.985	0.992	0.993	0.997
1057	6.682×10^{-12}	R	0.998	1.000	1	1	951	3.462×10^{-7}	R	0.983	0.991	0.991	0.996
1030	1.744×10^{-10}	R	0.995	0.999	0.999	1	942	7.095×10^{-7}	R	0.986	0.992	0.993	0.998
1027	2.445×10^{-10}	R	0.997	0.999	1	1	918	4.251×10^{-6}	R	0.962	0.976	0.976	0.986
1013	1.116×10^{-9}	R	0.992	0.998	0.998	1	900	1.458×10^{-5}	R	0.959	0.971	0.971	0.983
1000	4.219×10^{-9}	R	0.996	0.999	0.999	1	899	1.557×10^{-5}	R	0.935	0.952	0.953	0.967
1000	4.219×10^{-9}	R	0.997	0.999	1	1	879	5.502×10^{-5}	R	0.933	0.947	0.948	0.962
998	5.144×10^{-9}	R	0.991	0.996	0.997	1	873	7.875×10^{-5}	R	0.924	0.943	0.943	0.959

Table 3.8: RP for the JT test under H_1, with Case 2, $n = 20$, $J_{0.0497} = 722$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
79	0.023	R	0.412	0.460	0.460	0.490	59	0.361	NR	0.812	0.839	0.838	0.873
75	0.049	R	0.345	0.396	0.395	0.436	57	0.421	NR	0.824	0.850	0.851	0.878
74	0.058	NR	0.595	0.631	0.630	0.671	51	0.609	NR	0.852	0.880	0.881	0.907
72	0.080	NR	0.597	0.636	0.636	0.677	50	0.640	NR	0.949	0.961	0.962	0.974
70	0.107	NR	0.655	0.683	0.682	0.714	49	0.669	NR	0.939	0.955	0.955	0.968
67	0.159	NR	0.713	0.749	0.749	0.781	47	0.725	NR	0.950	0.966	0.967	0.978
66	0.180	NR	0.728	0.760	0.760	0.791	47	0.725	NR	0.927	0.944	0.944	0.957
62	0.276	NR	0.786	0.826	0.825	0.857	44	0.799	NR	0.894	0.919	0.919	0.938
60	0.331	NR	0.778	0.810	0.810	0.841	39	0.893	NR	0.951	0.966	0.965	0.982
60	0.331	NR	0.816	0.839	0.838	0.872	34	0.951	NR	0.971	0.983	0.983	0.992

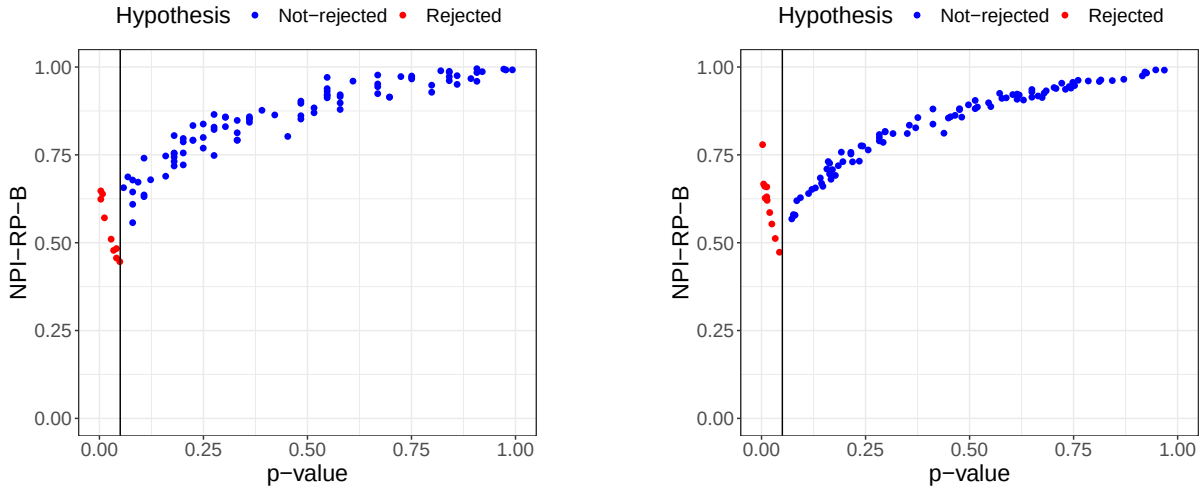
Table 3.9: RP for the JT test under H_1, with Case 3, $n = 6$, $J_{0.0490} = 75$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
771	0.010	R	0.583	0.632	0.632	0.665	609	0.454	NR	0.826	0.853	0.854	0.876
733	0.036	R	0.450	0.504	0.504	0.545	609	0.454	NR	0.844	0.868	0.868	0.890
724	0.047	R	0.419	0.469	0.469	0.515	593	0.540	NR	0.869	0.890	0.890	0.912
710	0.069	NR	0.537	0.574	0.576	0.622	589	0.562	NR	0.881	0.898	0.898	0.915
708	0.073	NR	0.550	0.581	0.581	0.626	574	0.640	NR	0.904	0.925	0.925	0.946
666	0.188	NR	0.686	0.717	0.717	0.751	551	0.748	NR	0.946	0.959	0.959	0.978
651	0.248	NR	0.770	0.797	0.799	0.827	540	0.793	NR	0.916	0.932	0.931	0.954
644	0.279	NR	0.768	0.800	0.802	0.826	525	0.846	NR	0.949	0.966	0.967	0.982
638	0.306	NR	0.835	0.856	0.858	0.880	509	0.892	NR	0.965	0.977	0.977	0.990
613	0.433	NR	0.829	0.848	0.848	0.875	405	0.996	NR	0.995	0.998	0.998	1

Table 3.10: RP for the JT test under H_1, with Case 3, $n = 20$, $J_{0.0497} = 722$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
80	0.019	R	0.457	0.492	0.493	0.528	55	0.484	NR	0.870	0.897	0.896	0.918
77	0.034	R	0.371	0.418	0.418	0.461	52	0.579	NR	0.880	0.898	0.897	0.923
74	0.058	NR	0.592	0.631	0.632	0.664	50	0.640	NR	0.940	0.952	0.951	0.966
73	0.068	NR	0.569	0.609	0.609	0.655	49	0.669	NR	0.942	0.958	0.958	0.969
72	0.080	NR	0.608	0.637	0.636	0.674	47	0.725	NR	0.946	0.962	0.962	0.977
69	0.123	NR	0.682	0.712	0.712	0.751	47	0.725	NR	0.919	0.936	0.937	0.953
65	0.202	NR	0.716	0.749	0.750	0.779	43	0.820	NR	0.976	0.985	0.986	0.993
65	0.202	NR	0.703	0.739	0.738	0.770	40	0.877	NR	0.941	0.958	0.957	0.973
61	0.303	NR	0.794	0.823	0.822	0.852	38	0.907	NR	0.967	0.978	0.978	0.988
57	0.421	NR	0.825	0.848	0.847	0.877	38	0.907	NR	0.975	0.984	0.984	0.994

Table 3.11: RP for the JT test under H_1, with Case 4, $n = 6$, $J_{0.0490} = 75$

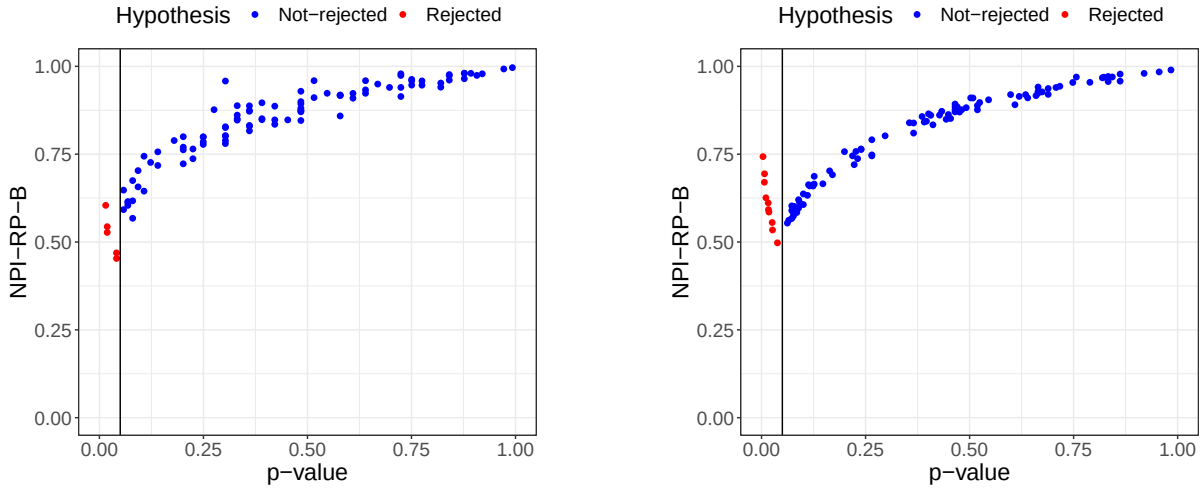


(a) $n = 6$ (b) $n = 20$

Figure 3.4: NPI-RP-B for the JT test under H_1, with Case 3, $\alpha = 0.05$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
797	0.004	R	0.662	0.709	0.711	0.749	660	0.211	NR	0.673	0.708	0.708	0.742
762	0.014	R	0.555	0.607	0.608	0.644	652	0.243	NR	0.757	0.787	0.789	0.819
746	0.024	R	0.494	0.551	0.552	0.582	648	0.261	NR	0.739	0.774	0.775	0.799
734	0.035	R	0.456	0.514	0.514	0.549	621	0.391	NR	0.814	0.836	0.835	0.859
732	0.037	R	0.471	0.516	0.515	0.545	619	0.402	NR	0.841	0.868	0.868	0.890
715	0.060	NR	0.534	0.569	0.570	0.620	599	0.508	NR	0.848	0.870	0.870	0.899
700	0.089	NR	0.567	0.609	0.609	0.651	571	0.655	NR	0.898	0.920	0.919	0.939
688	0.118	NR	0.598	0.637	0.637	0.680	548	0.761	NR	0.932	0.949	0.949	0.969
669	0.177	NR	0.713	0.746	0.746	0.788	473	0.958	NR	0.990	0.996	0.996	0.999
669	0.177	NR	0.713	0.746	0.746	0.788	423	0.992	NR	0.992	0.997	0.997	1

Table 3.12: RP for the JT test under H_1, with Case 4, $n = 20$, $J_{0.0497} = 722$



(a) $n = 6$ (b) $n = 20$

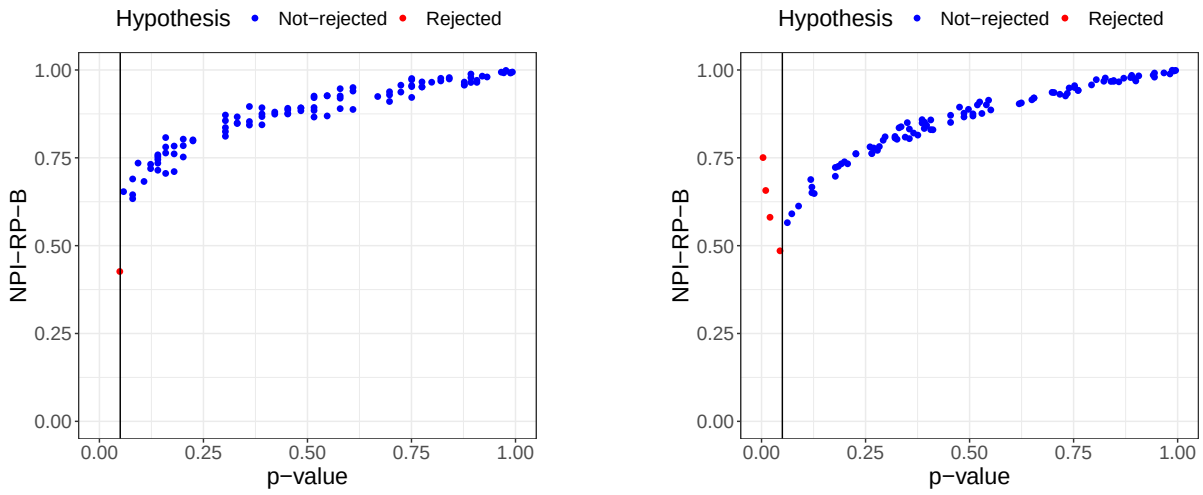
Figure 3.5: NPI-RP-B for the JT test under H_1, with Case 4, $\alpha = 0.05$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
76	0.041	R	0.354	0.401	0.402	0.441	47	0.725	NR	0.930	0.953	0.953	0.968
75	0.049	R	0.346	0.381	0.380	0.433	42	0.841	NR	0.948	0.965	0.965	0.978
74	0.058	NR	0.559	0.592	0.593	0.634	40	0.877	NR	0.953	0.964	0.964	0.979
73	0.068	NR	0.583	0.624	0.624	0.654	37	0.920	NR	0.977	0.987	0.987	0.995
71	0.093	NR	0.595	0.639	0.639	0.677	32	0.966	NR	0.976	0.985	0.985	0.993
65	0.202	NR	0.734	0.768	0.768	0.803	30	0.977	NR	0.980	0.989	0.990	0.995
64	0.225	NR	0.745	0.783	0.783	0.826	27	0.015	NR	0.980	0.988	0.987	0.995
60	0.331	NR	0.841	0.866	0.866	0.905	23	0.996	NR	0.988	0.994	0.994	0.998
55	0.484	NR	0.842	0.867	0.867	0.900	21	0.997	NR	0.993	0.997	0.998	1
50	0.640	NR	0.885	0.915	0.915	0.935	16	1.000	NR	0.997	1.000	1	1

Table 3.13: RP for the JT test, with Case 5, $n = 6$, $J_{0.0490} = 75$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
772	0.010	R	0.607	0.650	0.649	0.679	599	0.508	NR	0.863	0.884	0.883	0.905
741	0.028	R	0.506	0.543	0.543	0.575	580	0.609	NR	0.885	0.910	0.909	0.928
711	0.067	NR	0.547	0.583	0.580	0.621	559	0.712	NR	0.920	0.941	0.940	0.956
703	0.083	NR	0.577	0.611	0.613	0.646	522	0.856	NR	0.945	0.965	0.965	0.979
696	0.098	NR	0.570	0.609	0.609	0.662	511	0.887	NR	0.955	0.969	0.970	0.982
663	0.199	NR	0.701	0.735	0.735	0.775	511	0.887	NR	0.962	0.981	0.981	0.991
656	0.227	NR	0.702	0.735	0.734	0.778	505	0.902	NR	0.958	0.976	0.976	0.988
621	0.391	NR	0.809	0.840	0.841	0.865	490	0.933	NR	0.964	0.980	0.981	0.989
608	0.460	NR	0.810	0.841	0.840	0.866	469	0.963	NR	0.986	0.993	0.993	0.999
608	0.460	NR	0.865	0.890	0.891	0.913	449	0.980	NR	0.987	0.994	0.994	0.998

Table 3.14: RP for the JT test under H_0, with Case 5, $n = 20$, $J_{0.0497} = 722$

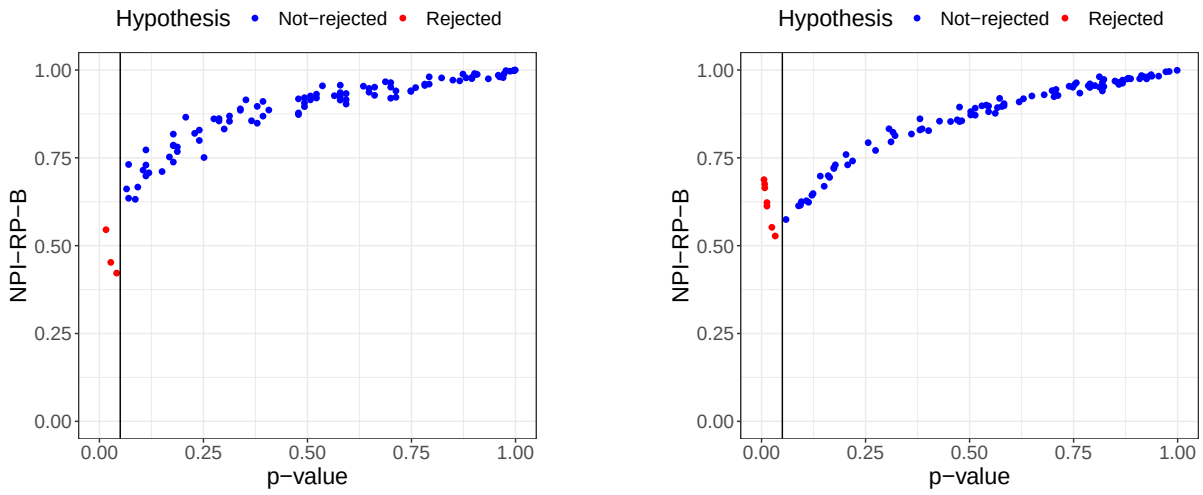


(a) $n = 6$ (b) $n = 20$

Figure 3.6: NPI-RP-B for the JT test under H_1, with Case 5, $\alpha = 0.05$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
227	0.045	R	0.360	0.405	0.407	0.447	165	0.713	NR	0.922	0.940	0.941	0.958
224	0.056	NR	0.611	0.647	0.649	0.683	160	0.771	NR	0.933	0.952	0.952	0.967
219	0.081	NR	0.630	0.662	0.662	0.702	156	0.813	NR	0.943	0.960	0.960	0.976
210	0.142	NR	0.667	0.701	0.702	0.743	150	0.866	NR	0.977	0.988	0.988	0.996
203	0.208	NR	0.733	0.766	0.766	0.807	150	0.866	NR	0.952	0.967	0.966	0.980
195	0.300	NR	0.773	0.803	0.804	0.831	147	0.888	NR	0.953	0.969	0.968	0.982
188	0.393	NR	0.842	0.866	0.866	0.896	147	0.888	NR	0.942	0.962	0.962	0.976
181	0.493	NR	0.838	0.866	0.865	0.895	131	0.965	NR	0.987	0.993	0.993	0.999
174	0.593	NR	0.933	0.955	0.955	0.972	128	0.973	NR	0.991	0.996	0.996	1
174	0.593	NR	0.922	0.939	0.938	0.957	119	0.988	NR	0.994	0.998	0.998	1

Table 3.15: RP for the JT test under H_0, with Case 6, $n = 6$, $J_{0.0484} = 226$



(a) $n = 6$ (b) $n = 20$

Figure 3.7: NPI-RP-B for the JT test under H_1, with Case 6, $\alpha = 0.05$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
2278	0.046	R	0.428	0.470	0.471	0.505	1935	0.655	NR	0.893	0.916	0.917	0.934
2261	0.057	NR	0.515	0.547	0.548	0.577	1902	0.725	NR	0.925	0.942	0.942	0.955
2228	0.083	NR	0.560	0.608	0.610	0.645	1892	0.745	NR	0.943	0.959	0.959	0.973
2195	0.119	NR	0.643	0.667	0.666	0.695	1850	0.820	NR	0.930	0.951	0.951	0.965
2157	0.171	NR	0.701	0.731	0.732	0.761	1814	0.871	NR	0.966	0.978	0.978	0.987
2148	0.185	NR	0.692	0.727	0.726	0.753	1736	0.946	NR	0.977	0.987	0.987	0.996
2114	0.245	NR	0.743	0.775	0.776	0.799	1660	0.981	NR	0.989	0.995	0.996	1
2093	0.287	NR	0.766	0.806	0.808	0.835	1609	0.992	NR	0.992	0.998	0.998	1
2071	0.334	NR	0.790	0.829	0.829	0.858	1556	0.997	NR	0.995	0.999	0.999	1
1988	0.530	NR	0.873	0.896	0.896	0.914	1397	1.000	NR	0.997	1.000	1	1

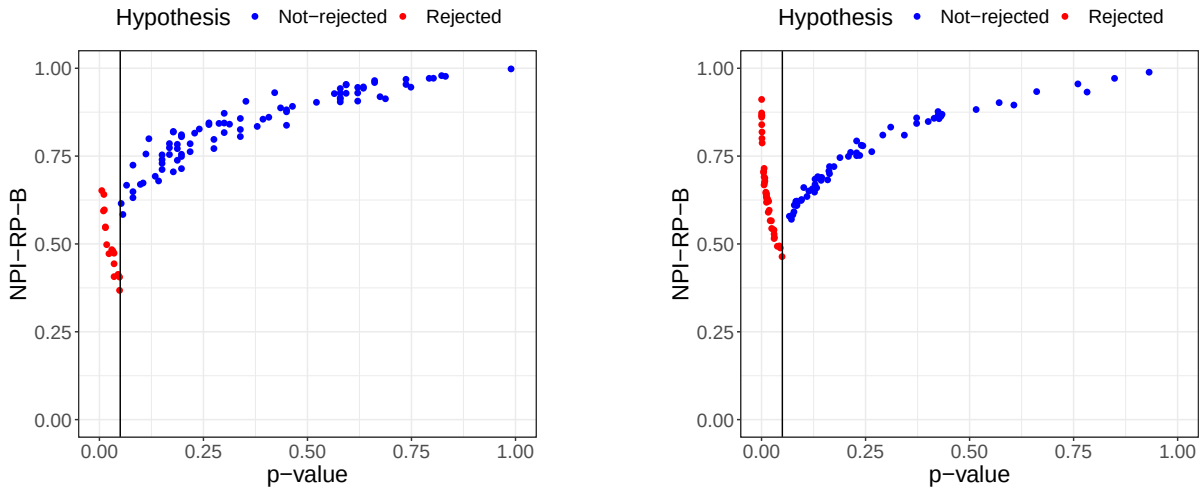
Table 3.16: RP for the JT test under H_0, with Case 6, $n = 20$, $J_{0.0499} = 2271$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
269	4.479×10^{-4}	R	0.834	0.858	0.859	0.877	218	0.086	NR	0.627	0.671	0.672	0.707
267	5.970×10^{-4}	R	0.851	0.882	0.881	0.907	214	0.112	NR	0.678	0.712	0.712	0.741
255	0.003	R	0.654	0.684	0.685	0.720	214	0.112	NR	0.651	0.686	0.684	0.728
248	0.006	R	0.528	0.574	0.575	0.612	209	0.151	NR	0.691	0.716	0.715	0.747
236	0.021	R	0.437	0.500	0.499	0.550	201	0.229	NR	0.799	0.831	0.832	0.857
233	0.027	R	0.389	0.430	0.431	0.478	191	0.352	NR	0.849	0.881	0.881	0.904
229	0.038	R	0.424	0.470	0.471	0.513	184	0.450	NR	0.837	0.867	0.868	0.900
225	0.052	NR	0.551	0.586	0.586	0.632	175	0.579	NR	0.883	0.908	0.908	0.931
225	0.052	NR	0.528	0.579	0.580	0.620	155	0.822	NR	0.976	0.986	0.986	0.994
222	0.065	NR	0.566	0600	0.602	0.633	147	0.888	NR	0.976	0.985	0.985	0.993

Table 3.17: RP for the JT test under H_1, with Case 7, $n = 6$, $J_{0.0484} = 226$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
2518	7.584×10^{-4}	R	0.815	0.837	0.837	0.862	2206	0.106	NR	0.623	0.659	0.659	0.692
2469	0.002	R	0.755	0.780	0.780	0.810	2190	0.125	NR	0.628	0.652	0.653	0.683
2457	0.003	R	0.711	0.737	0.737	0.768	2151	0.180	NR	0.682	0.715	0.716	0.747
2427	0.005	R	0.673	0.706	0.707	0.736	2072	0.332	NR	0.764	0.798	0.798	0.828
2375	0.011	R	0.596	0.633	0.632	0.666	2029	0.431	NR	0.814	0.854	0.854	0.877
2330	0.022	R	0.533	0.567	0.570	0.605	2003	0.494	NR	0.886	0.912	0.912	0.932
2308	0.031	R	0.501	0.538	0.540	0.570	1964	0.588	NR	0.896	0.914	0.915	0.933
2286	0.041	R	0.462	0.496	0.497	0.528	1923	0.681	NR	0.917	0.932	0.932	0.948
2280	0.044	R	0.438	0.483	0.484	0.515	1842	0.832	NR	0.953	0.968	0.968	0.980
2242	0.071	NR	0.535	0.581	0.581	0.615	1789	0.901	NR	0.969	0.981	0.981	0.990

Table 3.18: RP for the JT test under H_1, with Case 7, $n = 20$, $J_{0.0499} = 2271$



(a) $n = 6$ (b) $n = 20$

Figure 3.8: NPI-RP-B for the JT test under H_1, with Case 7, $\alpha = 0.05$

3.6 Concluding remarks

This chapter contributed to the development of NPI reproducibility by exploring the reproducibility probability for the Jonckheere-Terpstra test via the implementations of NPI-B. The test reproducibility is more naturally considered as a prediction problem than as an estimation problem. The NPI-B is explicitly predictive approach which considers future observations and is aligned well with the nature of test reproducibility. The use of the NPI-B to study RP avoids the complex calculations of the lower and upper NPI-RP, as well as it is a flexible approach to use when considering large sample sizes.

The NPI-RP-B method has been applied to a variety of scenarios via simulation studies. To sum up, the investigation in this chapter implies that when the the test statistics values are close to the test threshold the NPI-RP do not provide strong evidence in favour of the reproducibility of the test results, particularly, when H_0 is rejected more than when H_0 is not rejected. The reason for that is the presence of some sort of direction in the alternatives. Moving away from the threshold leads to large reproducibility probability estimates to about 90% in some cases which means that if the test is repeated in the future, there are 90% probability that the same conclusion would be reached.

There are many research challenges for the further development of NPI for reproducibility probability. For example, the reproducibility of the JT test in this chapter was investigated using the NPI-RP-B approach. However, deriving the JT test's exact lower and upper reproducibility probabilities is of interest for future work which may require developing some methods.

Further work entails exploring the calculation of NPI-RP estimates for the JT test via the the application of parametric predictive bootstrap methodology, introduced by Aldawsari [2]. The reproducibility for other ordered alternatives tests such as the Modified Jonckheere-Terpstra test [106] and the Page test [86] can be explored.

Chapter 4

Reproducibility of Umbrella Alternatives Tests

4.1 Introduction

In the one-way layout setting, the researchers are often concerned with detecting deviations from the null hypothesis that the location parameters are equal, indicating no group effect. Particular deviations of interest have included the general alternative (i.e., there is a group effect), the ordered alternative (i.e., there is monotone group effect), and the umbrella alternative (i.e., there is a monotone alternative that is subject to change in direction after reaching a peak). This chapter focuses on nonparametric tests for umbrella alternatives. Practical scenarios in which one would be concerned with detecting umbrella alternatives include experiments measuring responses to increasing drug dosage levels where an initial increasing effect culminating with a peak point (corresponding to the optimal dosage) and a decreasing effect afterwards.

Chapters 2 and 3, introduced the NPI for reproducibility probability for the general alternative tests and the ordered alternative test, respectively. This chapter contributes to the development of NPI for reproducibility by considering tests for the umbrella alternatives, namely the Mack-Wolfe (MW) test and the Esra-Fikri (EF) test. Utilising Sections 1.4 and 1.5, NPI approach to derive exact lower and upper reproducibility probabilities for the MW test and the EF test is introduced. However, exact NPI lower and upper reproducibility probabilities can only be computed for relatively small sample size. Calculating exact lower and upper probabilities for datasets with a large sample size is computationally challenging due to the increasing number of possible orderings of future observations, resulting in longer computational time. Marques et al. [75] encountered the same challenge while exploring the NPI reproducibility probability for reproducibility of likelihood ratio tests. To overcome this computational limitations, Marques and

Coolen [74] proposed the NPI sampling of orderings methodology (NPI-RP-SO). The NPI-RP-SO method provides approximation of the NPI lower and upper reproducibility probabilities. Another approach to address reproducibility probability for scenarios with large sample sizes is NPI bootstrap, which provides a point estimate for NPI reproducibility probabilities [18].

This chapter is organised as follows: Section 4.2 gives an overview of the classical tests for umbrella alternatives. In Section 4.3, the NPI approach is used to derive the exact lower and upper reproducibility probabilities for the MW test and the EF test for three groups. However, for more than three groups, computational difficulties prevent deriving the minimum and maximum values of the MW and the EF test statistics. Section 4.4 presents the NPI-RP using sampling of orderings method (NPI-RP-SO), to obtain approximations of the NPI lower and upper reproducibility probabilities for the MW test and the EF test for three groups. Section 4.5 introduces the methodology of the NPI bootstrap (NPI-RP-B), to estimate the reproducibility probability for the MW test and the EF test. In Section 4.6, illustrative examples are provided. NPI reproducibility probability for the MW test and the EF test is investigated via simulation in Section 4.7, considering the NPI-RP-SO and the NPI-RP-B approaches. Section 4.8 presents concluding remarks for this chapter.

4.2 Umbrella alternatives tests

The comparison of $k \geq 3$ groups in a one-way ANOVA setting include the situation that is the response variable may increase with the group level up to a certain point and then decrease [59]. This situation is common in many real problems, such as the effect of age on some variables that measure the physical capability such as muscle strength. Another example is the reaction to the increase of a drug dosage, which is increasing up to a certain point and then it decreases. This 'up-then-down' behavior is known in the literature as umbrella ordering [59]. The label umbrella was given to these alternatives by Mack and Wolfe in 1981 [71]. Let μ_i denote the location parameter for the ith population, $i = 1, 2, \ldots, k$. Several k-sample rank tests are introduced to test the following hypothesis,

$$H_0 : \mu_1 = \mu_2 = \ldots = \mu_k \tag{4.1}$$

against the umbrella alternatives

$$H_1 : \mu_1 \leq \mu_2 \leq \ldots \leq \mu_{p-1} \leq \mu_p \quad \text{and} \quad \mu_p \geq \mu_{p+1} \geq \ldots \geq \mu_k \tag{4.2}$$

for some $p \in \{1, 2, ..., k\}$, with at least one strict inequality. The umbrella alternative is said to have a peak at population p, where p can be either known or unknown [59]. Many of

these k-sample rank tests for testing umbrella alternatives are based on the Mann-Whitney test statistic [73], which has been the framework for many tests involving ordered alternatives [16, 45, 71, 76]. Moreover, umbrella alternatives test statistics are defined by combining the sums of Mann-Whitney statistics to the left and to the right of the peak group only and they do not include comparisons across the peak. Hettmansperger and Norton [56] proposed tests for testing the umbrella alternatives for both known peak and unknown peak and have pointed out that the absence of comparisons across peaks can cause some loss of efficiency. Other related methodologies and approaches constructing test statistics for umbrella alternatives were also presented in the literature, see e.g. [13, 22, 56, 72]. In this chapter, we consider the reproducibility probability for the Mack-Wolfe (MW) test [71] and the Esra-Fikri (EF) test [45].

4.2.1 Mack-Wolfe (MW) test

Mack and Wolfe [71] considered the umbrella alternatives with both known and unknown peak. The Mack-Wolfe statistic A_p for known peak p, uses the $p(p-1)/2$ Mann–Whitney counts, U_{uv}, for every pair of groups with $1 \leq u < v \leq p$, where U_{uv} is the number of observations from sample u that are smaller than the observations from sample v. The statistic A_p uses the $(k-p+1)(k-p)/2$ reverse Mann–Whitney counts U_{vu} for every pair of groups with $p \leq u < v \leq k$. Thus, the Mack–Wolfe peak-known statistic A_p is the sum of the Mann–Whitney counts to the left of the peak and the reverse Mann–Whitney counts to the right of the peak, as follows

$$A_p = \sum_{u=1}^{v-1} \sum_{v=2}^{p} U_{uv} + \sum_{u=p}^{v-1} \sum_{v=p+1}^{k} U_{vu} \tag{4.3}$$

The null hypothesis is rejected, at level of significance α, if and only if

$$A_p \geq A_{p,\alpha}$$

where $A_{p,\alpha}$ is the α-upper percentile for the null distribution of A_p, which can be computed using the function `cUmbrPK(`α, n, p`)` from the R package `NSM3` [53] and also can be found from tables in [20, 71].

For large sample sizes, and under the null hypothesis, the statistic A_p is asymptotically Normally distributed with the following mean and variance [59]:

$$E(A_p) = \frac{N_1^2 + N_2^2 - \sum_{i=1}^{k} n_i^2 - n_p^2}{4} \tag{4.4}$$

$$\sigma^2(A_p) = \frac{1}{72}\Bigg\{2(N_1^3 + N_2^3) + 3(N_1^2 + N_2^2) - \sum_{i=1}^{k} n_i^2(2n_i + 3) - n_p^2(2n_p + 3) + 12n_pN_1N_2 - 12n_p^2N\Bigg\} \tag{4.5}$$

where n_i is the size of sample i, $N_1 = \sum_{i=1}^{p} n_i$ and $N_2 = \sum_{i=p}^{k} n_i$. The observations in the peak group p are counted in both N_1 and N_2, thus $N = N_1 + N_2 - n_p$. The standardized version of A_p is

$$A'_p = \frac{A_p - E(A_p)}{\sigma(A_p)}$$

The Mack–Wolfe statistic for unknown peak p uses the sample data to estimate p, that is, we use the sample data to estimate which of the groups is most likely to correspond to the peak of the umbrella by calculating k combined samples Mann–Whitney statistics as follows:

$$U_{\cdot q} = \sum_{i \neq q} U_{iq}, \quad \text{for } q = 1, ..., k \tag{4.6}$$

where U_{iq} is the number of observations from the ith sample that are smaller than the observations from the qth sample. Then, under the null hypothesis, each $U_{\cdot q}$ is standardize as follows

$$E(U_{\cdot q}) = \frac{n_q(N - n_q)}{2} \tag{4.7}$$

$$\sigma^2(U_{\cdot q}) = \frac{n_q(N - n_q)(N + 1)}{12} \tag{4.8}$$

Thus, the standardized version of $U_{\cdot q}$ is given by

$$U'_{\cdot q} = \frac{U_{\cdot q} - E(U_{\cdot q})}{\sigma(U_{\cdot q})}, \quad q = 1, ..., k \tag{4.9}$$

Notice that, when the sample sizes of all groups are equal, the group with the largest $U_{\cdot q}$ value will also be the one with the largest $U'_{\cdot q}$ value. Let s be the number of groups that are tied for having the maximum $U'_{\cdot q}$ value and let D be the subset of $\{1, 2, ..., k\}$ that corresponds to the s groups tied for the maximum $U'_{\cdot q}$ value. The Mack–Wolfe peak unknown statistic is then given by

$$A'_{\hat{p}} = \frac{1}{s} \sum_{j \in D} \frac{A_j - E(A_j)}{\sigma(A_j)} \tag{4.10}$$

where A_j is the peak-known statistic with peak at the jth group, as given in Equation (4.3), and $E(A_j)$ and $\sigma(A_j)$ are given by Equations (4.4) and (4.5), respectively. The null hypothesis is rejected, at level of significance α, if and only if

$$A'_{\hat{p}} \geq A^*_{\hat{p},\alpha} \tag{4.11}$$

In most cases, $s = 1$ and $A'_{\hat{p}}$ is equal to the single standardized peak-known statistic A'_p. $A^*_{\hat{p},\alpha}$ is the upper α percentile for the null distribution of $A'_{\hat{p}}$ which can be computed using the function `cUmbrPU(`α, n`)` from the R package `NSM3` [53] and also found in tables [20, 71]

4.2.2 Esra-Fikri (EF) test

Esra and Fikri [45] proposed a modified version of the Mack-Wolfe test for umbrella alternatives, considering both known and unknown peak. For the case when the peak is known, the modified Mack-Wolfe statistic $\tilde{A}_p$ is the weighted sum of the Mann-Whitney counts to the left of the peak, $(v-u)U_{uv}$, and the reverse Mann-Whitney counts to the right of the peak, $(v-u)U_{vu}$. This modified test statistic gives weight 1 to Mann-Whitney statistics between adjacent groups [45]. The EF test statistic is given by

$$\tilde{A}_p = \sum_{u=1}^{p-1} \sum_{v=u+1}^{p} (v-u)U_{uv} + \sum_{u=p}^{k-1} \sum_{v=u+1}^{k} (v-u)U_{vu} \tag{4.12}$$

For balanced data, i.e. $n_1 = \ldots = n_k = n$, and under H_0, the EF statistic $\tilde{A}_p$ in Equation (4.12), is asymptotically Normally distributed with the following mean and variance:

$$E(\tilde{A}_p) = \frac{n^2}{2}\left[\binom{p+1}{3} + \binom{k-p+2}{3}\right] \tag{4.13}$$

$$\begin{aligned}\sigma^2(\tilde{A}_p) = {} & \frac{n^2p^2(p^2-1)(np+1)}{144} \\ & + \frac{n^2(k-p+1)^2[(k-p+1)^2-1][n(k-p+1)+1]}{144} \\ & + \frac{n^3p(p-1)(k-p)(k-p+1)}{24}\end{aligned} \tag{4.14}$$

The null hypothesis H_0 is rejected, at level of significance α, if and only if

$$\tilde{A}^*_p = \frac{\tilde{A}_p - E(\tilde{A}_p)}{\sigma(\tilde{A}_p)} \geq Z_\alpha \tag{4.15}$$

where Z_α is the $\alpha-$upper quantile of the standard Normal distribution.

For the case when the peak is unknown, the idea introduced in Section 4.2.1 for the Mack-Wolf test with unknown peak is used, such that the estimated peak $\hat{p}$ is the maximum of $(U'_{.1}, U'_{.2}, \ldots, U'_{.k})$ where $U'_{.q}$ is given in Equation (4.9) with $q = 1, 2, ..., k$. The standardized test

statistic for the unknown peak can be written as

$$\tilde{A^*}_{\hat{p}} = \frac{\tilde{A}_{\hat{p}} - E(\tilde{A}_{\hat{p}})}{\sigma(\tilde{A}_{\hat{p}})} \tag{4.16}$$

The null hypothesis is rejected, at level of significance α, if and only if

$$\tilde{A^*}_{\hat{p}} \geq Z_\alpha \tag{4.17}$$

4.3 NPI-RP-E for the Mack–Wolfe test

In this section, NPI-RP for the Mack-Wolfe test is introduced in term of the lower and upper reproducibility, denoted by $\underline{RP}$ and $\overline{RP}$, respectively. We consider the case of three independent groups X, Y and Z, with n_x observations from group X, n_y observations from group Y and n_z observations from group Z. Let $x_1 < ... < x_{n_x}$ be the ordered observed values of group X, these observations partition the real-line into $n_x + 1$ intervals $I_j^x = (x_{j-1}, x_j)$, $j = 1, ..., n_x + 1$. Let $y_1 < ... < y_{n_y}$ be the ordered observed values of group Y, these observations partition the real-line into $n_y + 1$ intervals $I_i^y = (y_{i-1}, y_i)$, $i = 1, ..., n_y + 1$. Let $z_1 < ... < z_{n_z}$ be the ordered observed values of group Z, these observations partition the real-line into n_z+1 intervals $I_k^z = (z_{k-1}, z_k)$, $k = 1, ..., n_z + 1$, and $x_0 = y_0 = z_0 = -\infty$ and $x_{n_x+1} = y_{n_y+1} = z_{n_z+1} = \infty$ for ease of notation. We also assume here that there are no tied observations. If tied observations occur, then these can be dealt with by a common method to break ties [58].

Let the number of future observations from groups X, Y, and Z be denoted by m_x, m_y and m_z, respectively. Here, we restrict attention to the case where the number of future observations is equal to the number of data observations ($m_x = n_x$, $m_y = n_y$ and $m_z = n_z$) which is considered a logical assumption in order to study reproducibility. There are $\binom{2n_x}{n_x}$ possible orderings of n_x future observations among the n_x data observations, where all possible orderings are equally likely. Similarly, there are $\binom{2n_y}{n_y}$ possible orderings of n_y future observations among the n_y data observations and there are $\binom{2n_z}{n_z}$ possible orderings of n_z future observations based on n_z data observations and all possible orderings are equally likely. We consider all $\binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$ combinations of these possible orderings, denoted by O_ℓ for $\ell = 1, 2, ..., \binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$.

For each combination of orderings O_ℓ, the corresponding Mack-Wolfe test statistic, given in Equation (4.3), denoted by $A_{p\ell}$. As the future observations are not precise, but only their number in each of the intervals of the partition created by the original data observations for their groups are known for a given ordering, we cannot calculate a precise value of $A_{p\ell}$ related to a specific combination of orderings, but we can derive the minimum and maximum possible values; these are denoted by $\underline{A_{p_\ell}}$ and $\overline{A_{p\ell}}$, respectively.

Let a specific ordering of n_x future observations among the n_x data observations be denoted by $(S_1^X, ..., S_{n_x+1}^X)$, with S_j^X non-negative integers with $\sum_{j=1}^{n_x+1} S_j^X = n_x$, as introduced in Section 1.4. Let a specific ordering of n_y future observations among the n_y data observations be denoted by $(S_1^Y, ..., S_{n_y+1}^Y)$, with S_i^Y non-negative integers with $\sum_{i=1}^{n_y+1} S_i^Y = n_y$. Let a specific ordering of n_z future observations among the n_z data observations be denoted by $(S_1^Z, ..., S_{n_z+1}^Z)$, with S_k^Z non-negative integers with $\sum_{k=1}^{n_z+1} S_k^Z = n_z$. In addition, let $j(i) = \max\{j : x_{(j)} < y_{(i)}\}$ for $i = 1, ..., n_y+1$ and $j = 0, 1, ..., n_x$, so $x_{(j(i))} < y_{(i)} < x_{(j(i)+1)}$ and the rank of $y_{(i)}$ in the combined ordered data from both groups X and Y is $i + j(i)$. Likewise, let $k(i) = \max\{k : z_{(k)} < y_{(i)}\}$ for $i = 1, ..., n_y + 1$ and $k = 0, 1, ..., n_z$, so $z_{(k(i))} < y_{(i)} < z_{(k(i)+1)}$ and the rank of $y_{(i)}$ in the combined ordered data from both groups Z and Y is $i+k(i)$. The minimum and the maximum values of $A_{p\ell}$, are as follows

$$\underline{A_{p_\ell}} = \sum_{i=1}^{n_y+1} S_i^Y \left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z \right] \tag{4.18}$$

$$\overline{A_{p_\ell}} = \sum_{i=1}^{n_y+1} S_i^Y \left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + \sum_{c=1}^{k_{(i)}-1} S_c^Z \right] \tag{4.19}$$

Note, the ℓ is omitted from the right hand side for simplicity of notation. The detailed justification for these results can be found in Appendix C.

The NPI lower (upper) reproducibility probability if the original test conclusion is rejection of H_0, is derived by counting the combinations for which $\underline{A_{p_\ell}} \geq A_{p,\alpha}$ ($\overline{A_{p\ell}} \geq A_{p,\alpha}$). Thus, the NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{h} \sum_{\ell=1}^{h} 1\{\underline{A_{p_\ell}} \geq A_{p,\alpha}\} \tag{4.20}$$

$$\overline{RP} = \frac{1}{h} \sum_{\ell=1}^{h} 1\{\overline{A_{p_\ell}} \geq A_{p,\alpha}\} \tag{4.21}$$

where $h = \binom{2n_x}{n_x}\binom{2n_y}{n_y}\binom{2n_z}{n_z}$ and $1\{A\}$ is an indicator function which is equal to 1 if the event A occurs and 0 otherwise. Similarly, if the conclusion of the original test is non-rejection of H_0 then the NPI lower (upper) reproducibility probability is derived by counting the combinations for which $\overline{A_{p_\ell}} < A_{p,\alpha}$ ($\underline{A_{p_\ell}} < A_{p,\alpha}$). Thus, the NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{h} \sum_{\ell=1}^{h} 1\{\overline{A_{p_\ell}} < A_{p,\alpha}\} \tag{4.22}$$

$$\overline{RP} = \frac{1}{h} \sum_{\ell=1}^{h} 1\{\underline{A_{p_\ell}} < A_{p,\alpha}\} \tag{4.23}$$

The exact NPI lower and upper reproducibility probabilities for the EF test for three groups with $p = 2$ which refers to the second group Y, are identical to the exact NPI-RP for the Mack-Wolfe (MW) test. This is because the EF test statistic for three groups X, Y and Z is the sum of two Mann-Whitney statistics between X and Y and between Z and Y, with a weight of one assigned to both of them, which is equal to the MW test statistic for three groups. Therefore, the NPI-RP-E method introduced in this section applies to both the EF test and the MW test. However, for large sample sizes, going through all possible orderings is computationally expensive. In this case, we can apply the method of sampling of orderings (NPI-RP-SO) and NPI bootstrap (NPI-RP-B) to derive approximations for the lower and upper reproducibility probabilities. The methodology of this section will be investigated using an artificial data sets of ranks.

4.4 NPI-RP-SO for the Mack-Wolfe test

The exact method to derive the NPI lower and upper reproducibility probabilities for the Mack-Wolfe test which was introduced in Section 4.3, is only computationally feasible for small sample sizes, as it considers all the orderings. For large sample sizes, the NPI-RP sampling of orderings (NPI-RP-SO) method is applied to obtain approximations for the lower and upper reproducibility probabilities. In the NPI-RP-SO approach, we randomly sample r^* orderings from all possible orderings of the future observations among the data observations per group [74, 75]. Then, apply Equations (4.18) and (4.19) on these sampled orderings to obtain the minimum and the maximum of the Mack-Wolfe test statistic, A_p. Here, all orderings sampled combinations will not be considered, instead consider each ordering sampled from one group with the corresponding ordering sampled from the other groups.

If the conclusion of the original test is rejection of H_0, so $A_p \geq A_{p,\alpha}$, then the NPI lower reproducibility probability using the NPI-RP-SO is obtained by counting the number of orderings for which $\underline{A_{p_\ell}} \geq A_{p,\alpha}$ and divided by the number of orderings sampled, r^*. The NPI upper reproducibility probability is obtained by counting the number of orderings for which $\overline{A_{p\ell}} \geq A_{p,\alpha}$, and divided it by, r^*:

$$\widehat{\underline{RP}} = \frac{1}{r^*} \sum_{\ell=1}^{r^*} \mathbf{1}\{\underline{A_{p_\ell}} \geq A_{p,\alpha}\} \tag{4.24}$$

$$\widehat{\overline{RP}} = \frac{1}{r^*} \sum_{\ell=1}^{r^*} \mathbf{1}\{\overline{A_{p\ell}} \geq A_{p,\alpha}\} \tag{4.25}$$

Similarly, if the original test conclusion is non-rejection of H_0, so $A_p < A_{p,\alpha}$, then the NPI lower

and upper reproducibility probabilities are given by

$$\widehat{\underline{RP}} = \frac{1}{r^*}\sum_{\ell=1}^{r^*} \mathbf{1}\{\overline{A_{p_\ell}} < A_{p,\alpha}\} \tag{4.26}$$

$$\widehat{\overline{RP}} = \frac{1}{r^*}\sum_{\ell=1}^{r^*} \mathbf{1}\{\underline{A_{p_\ell}} < A_{p,\alpha}\} \tag{4.27}$$

The NPI-RP-SO method will be illustrated via simulations and examples with data sets from the literature to investigate the NPI-RP for the MW test and EF test for three groups.

4.5 NPI-RP-B for the MW test and the EF test

In the previous section, the NPI-RP-SO is introduced to study the NPI-RP for the MW test and the EF test for large sample sizes. However, the NPI-RP-SO may not always be possible to use, as the application of NPI-RP-SO requires deriving the exact NPI lower and upper reproducibility probabilities, which could be challenging for some some test statistics. The NPI-RP-B method, introduced in 1.5.3, can be apply as an alternative method to approximate the reproducibility probability for the Mack-Wolfe (MW) test and the Esra and Fikri (EF) test. NPI-RP-B uses the point estimate to present the NPI reproducibility probability instead of lower and upper reproducibility probabilities. As mentioned earlier in this chapter, with large sample sizes, computational issues prevent the exact NPI reproducibility probability approach, introduced in Section 4.3. Moreover, computational difficulties prevent deriving the minimum and maximum for the MW and the EF test statistics for more than three groups. This can be resolved by using the NPI-RP-B method, as it avoids the complex calculations required by the exact NPI reproducibility probability approach.

Algorithm 2 introduced in Section 1.5.3, is applied. The inputs into Algorithm 1 are the k original samples, their corresponding sample sizes, the number of runs T and the number of bootstrapped samples per run B. Summary statistics including the minimum, mean, median, maximum, of $RP_1, RP_2, ..., RP_T$ were calculated. The mean value of the outcomes is the reproducibility probability estimate, and is referred to as NPI-RP-B value. In this chapter, Algorithm 2, will be implemented with both finite and infinite intervals, using Approaches I and II, introduced in Section 2.3. In Approach I, the lower limit is taken to be smallest value of the group minus the maximal distance between consecutive points, and the upper limit is taken to be equal to largest value of the group plus the maximal distance between consecutive points. Approach II involves assuming the tail of a Normal distribution for real-valued data and the tail of an Exponential distribution for non-negative real-valued data. It is important

to emphasize that the bootstrap samples have the same size as the original sample. Section 4.7 presents the results of simulation studies for different scenarios, such as simulation under H_0 and under H_1, with varying sample sizes and number of groups.

4.6 Examples

This section studies the reproducibility probability for the MW test and the EF test for three groups. In example 4.1, artificial data sets of ranks are used to investigate reproducibility probability using the NPI-RP-E, NPI-RP-SO and NPI-RP-B approaches and the results are compared. In example 4.2, the NPI-RP-SO approach is considered for large data set from the literature with equal sample sizes. The NPI-RP-SO approach is applied for large data set with unequal sample sizes in Example 4.3.

Example 4.1. This example investigates the reproducibility probability for the MW test and the EF test for $k = 3$ groups X, Y and Z by applying the NPI-RP-E approach, introduced in Section 4.3. Then, a comparison of the three methods, NPI-RP-E, NPI-RP-SO and NPI-RP-B is carried out investigate whether or not the NPI-RP-B method tends to provide a value within the lower and upper NPI-RP-E and NPI-RP-SO. Artificial data sets of ranks with equal samples sizes $n_x = n_y = n_z = 3$, and $n_x = n_y = n_z = 5$ are considered. The hypothesis of interest is $H_0 : \mu_x = \mu_y = \mu_z$ against $H_1 : \mu_x \leq \mu_y \geq \mu_z$, that is $p = 2$ which refers to the second group Y, at the level of significance $\alpha = 0.05$.

For the MW test with $n_x = n_y = n_z = 3$ and $n_x = n_y = n_z = 5$, in Tables 4.1 and 4.2, due to the discrete nature of the test statistic the nominal levels are 0.0476 and 0.0496, respectively. Accordingly, the test decision rule for the MW test is to reject the null hypothesis if the test statistic A_p is greater than or equal to $A_{2,0.0476} = 16$ and $A_{2,0.0496} = 39$, respectively. For the EF test, the test decision rule is to reject the null hypothesis if the test statistic $\tilde{A}_p^*$ is greater than or equal to $Z_{0.05} = 1.645$. Throughout this thesis, the original test conclusion is denoted by R when the null hypothesis is rejected and NR when the null hypothesis is not rejected, and the values in the tables are rounded to three decimal digits while precise value 1 is presented without additional decimals, so the values 1.000 are actually less than 1 but rounded up. The NPI-RP results presented in Tables 4.1 and 4.2 are identical for both the MW test and the EF test because we have three groups and $p = 2$. Therefore, the following discussion applies to both tests.

In order to calculate the exact lower and upper NPI-RP for the MW test and EF test for three groups with $n_x = n_y = n_z = 3$, there are $\binom{6}{3}$=20 possible orderings of 3 future

Ranks			Test conclusion				NPI-RP-E		NPI-RP-B				NPI-RP-SO	
X	Y	Z	A_p	$\tilde{A}_p^*$	p-value	H_0	$\underline{RP}$	$\overline{RP}$	Min	Mean	Median	Max	$\widehat{\underline{RP}}$	$\widehat{\overline{RP}}$
1,2,3	7,8,9	4,5,6	18	2.324	0.010	R	0.125	1	0.955	0.971	0.971	0.983	0.114	1
2,3,4	7,8,9	1,5,6	18	2.324	0.010	R	0.125	1	0.760	0.789	0.790	0.822	0.123	1
1,2,3	6,8,9	4,5,7	17	2.066	0.019	R	0.106	0.930	0.645	0.676	0.676	0.720	0.107	0.934
1,2,3	5,8,9	4,6,7	16	1.807	0.035	R	0.081	0.825	0.441	0.478	0.476	0.516	0.089	0.830
1,2,7	5,8,9	3,4,6	16	1.807	0.035	R	0.086	0.832	0.369	0.398	0.397	0.429	0.089	0.824
1,2,3	6,7,9	4,5,8	16	1.807	0.035	R	0.081	0.825	0.459	0.494	0.495	0.536	0.078	0.829
1,2,3	4,8,9	5,6,7	15	1.549	0.061	NR	0.318	0.950	0.591	0.639	0.640	0.676	0.314	0.953
2,3,4	5,7,9	1,6,8	15	1.549	0.061	NR	0.273	0.939	0.603	0.645	0.646	0.684	0.275	0.930
4,6,7	3,8,9	1,2,5	14	1.291	0.098	NR	0.386	0.950	0.656	0.690	0.691	0.722	0.414	0.953
4,5,6	1,8,9	2,3,7	12	0.775	0.219	NR	0.476	0.950	0.713	0.753	0.754	0.784	0.473	0.952
1,4,8	3,5,9	2,6,7	11	0.516	0.303	NR	0.578	0.977	0.826	0.865	0.866	0.888	0.566	0.977
1,2,3	4,5,6	7,8,9	9	0.000	0.500	NR	0.790	1	0.997	0.999	1	1	0.790	1
1,2,8	4,5,6	3,7,9	9	0.000	0.500	NR	0.720	0.995	0.944	0.958	0.958	0.973	0.715	0.997
1,3,4	2,5,6	7,8,9	7	-0.516	0.697	NR	0.833	1	0.986	0.993	0.993	1	0.824	1
1,2,6	3,4,5	7,8,9	6	-0.775	0.781	NR	0.855	1	1	1	1	1	0.846	1
1,2,9	3,4,5	6,7,8	6	-0.775	0.781	NR	0.855	1	0.998	1.000	1	1	0.848	1
5,3,9	1,2,8	7,4,6	5	-1.033	0.849	NR	0.814	0.995	0.926	0.948	0.947	0.966	0.818	0.995
1,4,5	2,3,6	7,8,9	5	-1.033	0.849	NR	0.870	1	0.987	0.993	0.993	1	0.872	1
1,4,7	2,3,5	6,8,9	4	-1.291	0.902	NR	0.889	1	0.996	0.999	0.999	1	0.892	1
4,5,6	1,2,3	7,8,9	0	-2.324	0.990	NR	0.933	1	1	1	1	1	0.932	1

Table 4.1: RP for the MW test and the EF test, with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, $n_x = n_y = n_z = 3$, $A_{2,0.0476} = 16$, $Z_{0.05} = 1.645$

observations among 3 data observations per group, so each $\underline{RP}$ and $\overline{RP}$ value is based on $\binom{6}{3}\binom{6}{3}\binom{6}{3} = 8000$ orderings combinations. In the NPI approach, with $n_x = n_y = n_z = 5$, there are $\binom{10}{5}$ possible orderings of 5 future observations among 5 data observations per group, and all $\binom{10}{5}\binom{10}{5}\binom{10}{5} = 1.600 \times 10^{17}$ orderings combinations are considered.

The NPI-RP results in Tables 4.1 and 4.2 show that $\underline{RP}$ is substantially less than 0.5 for several cases, and the $\underline{RP}$ value is low close to the test threshold and it is lower for the cases when the null hypothesis is rejected than for those not. This is because there exists some sort of direction in the alternative hypothesis. This implies that data, typically with test statistic is close to the threshold value and H_0 is rejected, do not provide strong evidence in favour of reproducibility of the test results.

For the exact lower reproducibility probability, H_0 is only rejected for the future samples if all the future Y ranks not in the first interval for Y i.e (greater than the smallest observed Y rank), and all future X ranks not in the last interval for X (smaller than the largest observed

Ranks			Test conclusion				NPI-RP-E		NPI-RP-B				NPI-RP-SO	
X	Y	Z	A_p	$\tilde{A}_p^*$	p-value	H_0	$\underline{RP}$	$\overline{RP}$	Min	Mean	Median	Max	$\widehat{\underline{RP}}$	$\widehat{\overline{RP}}$
1,2,3,4,5	11,12,13,14,15	6,7,8,9,10	50	3.062	0.001	R	0.441	1	0.997	0.999	0.999	1	0.443	1
1,2,6,7,8	11,12,13,14,15	3,4,5,9,10	50	3.062	0.001	R	0.441	1	0.974	0.985	0.985	0.997	0.443	1
1,2,3,4,5	10,12,13,14,15	6,7,8,9,11	49	2.939	0.002	R	0.402	0.997	0.966	0.977	0.977	0.987	0.406	0.999
1,2,3,4,5	9,12,13,14,15	6,7,8,10,11	48	2.817	0.002	R	0.367	0.988	0.916	0.939	0.939	0.957	0.367	0.991
1,2,3,4,5	10,11,12,13,14	6,7,8,9,15	45	2.450	0.007	R	0.300	0.932	0.805	0.832	0.832	0.860	0.300	0.927
1,2,3,4,5	9,10,11,13,15	6,7,8,12,14	43	2.205	0.014	R	0.224	0.884	0.695	0.730	0.730	0.767	0.223	0.878
1,2,3,4,5	8,9,10,14,15	6,7,11,12,13	41	1.960	0.025	R	0.172	0.807	0.536	0.586	0.586	0.627	0.174	0.812
1,2,3,4,6	5,11,12,13,14	7,8,9,10,15	40	1.837	0.033	R	0.178	0.775	0.434	0.474	0.474	0.509	0.174	0.784
1,2,3,4,15	5,10,12,13,14	6,7,8,9,11	39	1.715	0.043	R	0.161	0.754	0.393	0.434	0.433	0.469	0.157	0.764
1,3,5,6,14	7,10,11,12,13	2,4,8,9,15	38	1.592	0.056	NR	0.284	0.858	0.571	0.602	0.600	0.648	0.278	0.853
1,3,5,7,14	6,10,11,12,13	2,4,8,9,15	37	1.470	0.071	NR	0.322	0.872	0.617	0.653	0.653	0.698	0.322	0.865
1,2,3,7,11	6,8,9,12,15	4,5,10,13,14	35	1.225	0.110	NR	0.401	0.915	0.663	0.699	0.701	0.735	0.401	0.917
1,2,3,4,5	7,8,9,10,14	6,11,12,13,15	33	0.980	0.164	NR	0.521	0.957	0.732	0.763	0.763	0.794	0.498	0.954
1,2,3,4,5	6,7,8,9,10	11,12,13,14,15	25	0.000	0.500	NR	0.821	1	0.998	1.000	1	1	0.799	1
1,2,3,6,7	4,5,8,9,10	11,12,13,14,15	21	-0.490	0.688	NR	0.866	1	0.993	0.998	0.998	1	0.850	1
1,2,10,14,15	3,4,5,9,12	6,7,8,11,13	18	-0.857	0.804	NR	0.853	0.996	0.956	0.969	0.969	0.982	0.843	0.994
1,12,13,14,15	2,3,4,5,11	6,7,8,9,10	10	-1.837	0.967	NR	0.933	1.000	0.981	0.989	0.989	0.997	0.930	1
1,6,7,11,12	2,3,4,5,9	8,10,13,14,15	8	-2.082	0.981	NR	0.950	1.000	0.995	0.999	0.999	1	0.946	1
4,7,8,9,10	1,2,3,5,6	11,12,13,14,15	2	-2.817	0.998	NR	0.969	1	1	1	1	1	0.969	1
6,7,8,9,10	1,2,3,4,5	11,12,13,14,15	0	-3.062	0.999	NR	0.972	1	1	1	1	1	0.973	1

Table 4.2: RP for the MW test and the EF test, with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, $n_x = n_y = n_z = 5$, $\alpha = 0.05$, $A_{2,0.0496} = 39$, $Z_{0.05} = 1.645$

X rank), and all future Z ranks not in the last interval for Z (smaller than the largest observed Z rank). Thus, for the case in the first row in Table 4.1 with $\underline{RP} = 0.125$, the future Y ranks all greater than 7 and the future X ranks all less than 3, and the future Z ranks all less than 6. These individual events happen with probability 0.5 in the NPI framework, with the independence between the three groups leading to the lower reproducibility probability $0.5 \times 0.5 \times 0.5 = 0.125$. So, there are $\binom{5}{3}$ orderings for the future X ranks. Likewise, there are $\binom{5}{3}$ orderings for the future Y ranks and $\binom{5}{3}$ orderings for the future Z ranks, which is in total $\binom{5}{3}\binom{5}{3}\binom{5}{3} = 1000$ orderings combinations out of the 8000 orderings combinations.

If one of the future X ranks is larger than the largest observed X rank and one of the future Y ranks is smaller than the smallest observed Y rank, or one of the future Z ranks is larger than the largest observed Z rank, then for all cases will not reject H_0 because those intervals are unbounded and a future X rank could be larger than a future Y rank or a future Z rank could be larger than Y. Thus, for the extreme case in the last row in Table 4.1 with $\underline{RP} = 0.933$, all the future Y ranks greater than 3, and all the future X ranks smaller than 4, and all the future Z ranks smaller than 7.

Tables 4.1 and 4.2 also show that in most cases where different ranks per sample lead to the

same value of the test statistic, the values of NPI-RP are differ for each case, so they depend on the actual ranks per sample and not just on the value of the test statistic (the reported two cases with $A_p = 6$ in Table 4.1 are an exception, which is just due to the same numbers of combinations for which the test result happens to be repeated, it is not a general property).

In Table 4.1, the NPI lower and upper reproducibility probabilities are very imprecise due to the small number of data observations per group, that is there are big differences between the corresponding NPI lower and upper reproducibility probabilities. With larger sample sizes, as in Table 4.2, imprecision tends to be smaller, reflecting the large amount of information. On the other hand, for reproducibility close to 1, the imprecision is close to 0, whereas for low reproducibility, the imprecision is larger than for high reproducibility. The NPI-RP approach is strongly based on the data, and as such imprecision tends to be less with larger sample size.

For larger sample sizes, going through all combinations becomes quickly computationally infeasible. For example, for $n_x = n_y = n_z = 7$ the NPI-RP-E approach requires going through $\binom{14}{7}^3 = (3432)^3 = 4.024 \times 10^{10}$ orderings combinations of the future observations among the data observations to derive the values of $\underline{RP}$ and $\overline{RP}$. Thus, the application of other computational methods within NPI framework is required. These methods are NPI-RP-SO and NPI-RP-B which give approximate values for the NPI reproducibility probability.

The NPI-RP-SO is considered with $r^* = 2000$ orderings sampled. The NPI-RP-B method is applied using Algorithm 2 with Approach I, and $B = 1000$ and $T = 100$. Summary statistics including the minimum, mean, median, maximum, of RP_i, $i = 1, 2, \ldots, T$, are computed. Then, we examine whether the NPI-RP-B estimates are between the corresponding lower and upper NPI-RP-E and NPI-RP-SO. Based on the results presented in Tables 4.1 and 4.2, it can be inferred that, 100% of NPI-RP-B estimates are included in the bounds derived by the NPI-RP-E and the NPI-RP-SO methods. This is a logical finding that would always be expected due to the construction of the NPI lower and upper probabilities with no assumptions of probability masses assigned to intervals between two consecutive observations. This result may not hold in some rare cases, due to the randomness of the bootstrap inferences. These results give a good impression of NPI-RP-B because they show that these values are consistent with the bounds of NPI-RP-E and NPI-RP-SO. Similar results are observed by BinHimd [18] in her study of reproducibility probability for the one sample signed rank test and the Wilcoxon-Mann-Whitney test, where 100% of NPI-RP-B are located within the bound of NPI-RP-E. In his investigation of reproducibility probability for the likelihood ratio test, Aldawsari [2] found that 88% of NPI-RP-B values are included in the bounds of NPI-RP-SO, and this ratio can be considered good because it represents the most values.

	Data										Mean	Std. dev
Sales	343	495	602	666	796	813	894	920	960	1499	798.8	315.637
Production	126	156	216	291	345	488	516	542	546	1362	458.8	355.422
Research and Development	391	450	472	496	609	645	705	763	910	1309	675	273.985

Table 4.3: Telephone communications data, for Example 4.2

Cases	X	Y	Z	A_p	$\tilde{A}_p^*$	H_0
Case 1	Production	Sales	Research and Development	148	2.112	R
Case 2	Sales	Research and Development	Production	110	0.440	NR
Case 3	Sales	Production	Research and Development	42	-2.552	NR

Table 4.4: Telephone communications data, three cases, for Example 4.2

Example 4.2. This example considers the NPI-RP-SO method, introduced in Section 4.4, for the MW test and the EF test, using the Telephone Communications data set given in Table 4.3 [67], where a firm aims to improve the cost effectiveness of its communications. Ten home office executives were randomly selected from the Sales, Production and Research and Development departments to take part in the study.

We test the null hypothesis $H_0 : \mu_x = \mu_y = \mu_z$ against $H_1 : \mu_x \leq \mu_y \geq \mu_z$, that is $p = 2$ which refers to the second group Y. This test is performed with the level of significance $\alpha = 0.05$, this leads to the threshold $A_{2,0.0498} = 138$, so the null hypothesis is rejected if $A_p \geq 138$. For the EF test, the null hypothesis is rejected if $\tilde{A}_p^* \geq 1.645$. Three different cases are considered in this example, these are summarised in Table 4.4. As we have three groups with equal sample sizes, the NPI-RP results are identical for the EF test and the MW test. Thus, in this example, the NPI-RP results for the EF test will be omitted except the original test conclusion as in Table 4.4.

For the first case, the Mack-Wolfe test is applied to the data with Production as group X, sales as group Y, and Research and Development as group Z. This leads to the original test value $A_p = 148$ which is greater than 138, so, the test conclusion is rejection of the null hypothesis. For the second case, Sales is considered as group X, Research and Development as group Y, and Production as group Z. This leads to $A_p = 110 < 138$, so the null hypothesis is not rejected. For the third case, Sales is group X, Production is group Y, and Research and Development is group Z. This leads to $A_p = 42 < 138$, so the null hypothesis is not rejected.

To obtain the exact lower and upper reproducibility probabilities, there are $\binom{20}{10}\binom{20}{10}\binom{20}{10} = 6.307 \times 10^{15}$ orderings combinations to be considered, which is too large for the exact lower and upper reproducibility probabilities to be computed. Hence, the sampling of orderings method is

Case 1: Sales is the peak(Y)				
r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.350	(0.257, 0.443)	0.830	(0.756, 0.904)
500	0.284	(0.244, 0.324)	0.794	(0.759, 0.829)
1,000	0.364	(0.334, 0.394)	0.800	(0.775, 0.825)
5,000	0.331	(0.318, 0.344)	0.807	(0.796, 0.818)
10,000	0.327	(0.318, 0.336)	0.801	(0.793, 0.809)
50,000	0.320	(0.316, 0.324)	0.805	(0.802, 0.808)
100,000	0.322	(0.319, 0.325)	0.803	(0.801, 0.805)
150,000	0.320	(0.318, 0.322)	0.807	(0.805, 0.809)
Case 2: Research and development is the peak(Y)				
r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.650	(0.557, 0.743)	0.950	(0.907, 0.993)
500	0.680	(0.639, 0.721)	0.958	(0.940, 0.976)
1,000	0.622	(0.592, 0.652)	0.945	(0.931, 0.959)
5,000	0.656	(0.643, 0.669)	0.950	(0.944, 0.956)
10,000	0.658	(0.649, 0.667)	0.954	(0.950, 0.958)
50,000	0.661	(0.657, 0.665)	0.953	(0.951, 0.955)
100,000	0.663	(0.660, 0.666)	0.954	(0.953, 0.955)
150,000	0.664	(0.662, 0.666)	0.955	(0.954, 0.956)
Case 3: Production is the peak(Y)				
r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.970	(0.915, 0.994)	1	(0.964, 1)
500	0.978	(0.965, 0.991)	1	(0.993, 1)
1,000	0.976	(0.967, 0.985)	1	(0.996, 1)
5,000	0.976	(0.972, 0.980)	0.999	(0.999, 1)
10,000	0.980	(0.977, 0.983)	0.999	(1.000, 1)
50,000	0.979	(0.978, 0.980)	0.999	(1.000, 1)
100,000	0.979	(0.978, 0.980)	0.999	(1.000, 1)
150,000	0.978	(0.977, 0.979)	0.999	(1.000, 1)

Table 4.5: NPI-RP-SO for the MW test and the EF test with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$, for Example 4.2

applied using a random sample of orderings r^* from each group. Then, we use these randomly selected orderings of future data observations to compute the minimum and maximum for the Mack-Wolfe test statistic in Equations (4.18) and (4.19), to find approximate values of the NPI lower and upper reproducibility probabilities. The corresponding 95% confidence intervals (CI) is computed for $\widehat{\underline{RP}}$ and $\widehat{\overline{RP}}$. The confidence intervals are calculated using the standard result based on the Normal approximation in Equation (1.10). When computing the Normal approximation confidence interval, for some cases where $\widehat{\underline{RP}}$ is close to 0 or $\widehat{\overline{RP}}$ is close to 1, the lower bound of the CI can be less than 0 or the upper bound greater than 1. Thus, the exact $(1-\alpha)100\%$ confidence interval in Equations (1.11) and (1.12) is used, as explained in details in Section 1.5.2.

For Case 1 in Table 4.5, the NPI lower reproducibility probability is low because the test statistic value $A_p = 148$ is close to the threshold value 138 and the null hypothesis is rejected in the original test. For Case 3, the $\widehat{\underline{RP}}$ value is large because the test statistic $A_p = 42$ is away from the threshold 138. From this Table, it can be inferred that the difference between NPI-RP estimates with increasing r^* is in the second decimal place, which is not very notable. Therefore, it can be concluded that reasonable approximations of the NPI lower and upper reproducibility probabilities for the MW test and EF test, can be obtained by considering the number orderings sampled equal or greater than 10,000 which is a quite small number when compared with the number of all possible orderings. The first application of NPI-RP-SO for test reproducibility, carried out by Marques et al. [75] for the likelihood ratio test, suggests that the number of orderings sampled should be at least 2000 to achieve reasonable results. NPI-RP-SO provides a computationally efficient way to obtain the values of lower and upper reproducibility probabilities.

Example 4.3. This example illustrates the NPI-RP-SO method for the MW test with large sample size data using the `"LengthWeightData"` in the R package `StatCharrms`. The data set, as given in Table 4.6, contains records of fish that were exposed to different levels of chemical concentration. In this example, we consider the weight variable with three levels of concentrations, C0, C6 and C13, with sample sizes of 28, 29 and 30, respectively, to obtain approximations for the NPI lower and upper reproducibility probabilities for the MW test. There are a few repeated observations in the data set, so we added a small amount to make these observations distinct. The NPI-RP for the EF test is not provided in this example because the EF test is proposed for data sets with equal sample sizes.

We test the hypothesis $H_0 : \mu_x = \mu_y = \mu_z$ against $H_1 : \mu_x \leq \mu_y \geq \mu_z$, that is $p = 2$ which refers to the second group Y, at level of significance $\alpha = 0.05$. Three different cases are

Concentrations	Weight										Mean	Std. dev
C0	45	50	56	60	62	65	71	74	76	81	140.790	148.038
	84	86	89	89	89	91	94	95	101	102		
	110	117	141	208	307	377	381	741				
C6	57	65	72	79	80	81	86	86	94	95	149.037	112.160
	96	98	103	110	112	113	114	114	118	119		
	122	131	136	205	268	277	371	393	527			
C13	52	59	65	76	83	84	98	101	104	104	169.168	127.979
	104	108	111	112	117	119	121	124	130	133		
	137	144	162	166	341	372	373	428	432	515		

Table 4.6: `LenghtWeightData`, for Example 4.3

Cases	X	Y	Z	A_p	H_0
Case 1	C0	C13	C6	1071	R
Case 2	C0	C6	C13	892	NR
Case 3	C13	C0	C6	559	NR

Table 4.7: `LenghtWeightData`, three cases, for Example 4.3

considered in this example, these are presented in Table 4.7. Applying the NPI approach for real-valued observations, there there are $\binom{28+28}{28}$ orderings of 28 future observations among 28 data observations from C0, there are $\binom{29+29}{29}$ orderings of 29 future observations among 29 data observations from C6, and there there are $\binom{30+30}{30}$ orderings of 30 future observations among 30 data observations from C13. It is unfeasible to go through the combinations of these large numbers of orderings. We sampled different numbers of orderings to explore the performance of the NPI-RP-SO for the Mack-Wolfe test, as shown in Tables 4.8.

For Case 1, the Mack-Wolfe test is applied to the data with C0 is X group, C13 is Y and the peak group, and C6 is Z group. This leads to $A_p = 1071$, so, the original data lead to rejection of H_0 since the test statistic value $A_p = 1071$ is greater than threshold $A_{2,0.0498} = 1040$. For Case 2, the Mack-Wolfe test is performed with C0 is X group, C6 is group Y and the peak group, and C13 is group Z. This leads to $A_p = 892$ which is less than the test threshold $A_{2,0.0493} = 1025$. So, the null hypothesis is not rejected. For Case 3, we apply the Mack-Wolfe test with C13 is X group, C0 is group Y and the peak group, and C6 is group Z. This leads $A_p = 559$ which is less than the test threshold $A_{2,0.0497} = 1008$. Thus, the null hypothesis is not rejected. Typically with test statistic values not close to the threshold, the data provide a

Case 1				
r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.350	(0.257, 0.443)	0.720	(0.632, 0.808)
500	0.406	(0.363, 0.449)	0.704	(0.664, 0.744)
1,000	0.417	(0.386, 0.448)	0.726	(0.698, 0.754)
5,000	0.405	(0.391, 0.419)	0.713	(0.700, 0.726)
10,000	0.409	(0.399, 0.419)	0.707	(0.698, 0.716)
50,000	0.412	(0.408, 0.416)	0.697	(0.693, 0.701)
100,000	0.410	(0.407, 0.413)	0.700	(0.697, 0.703)
150,000	0.409	(0.407, 0.411)	0.699	(0.697, 0.701)
Case 2				
r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.620	(0.525, 0.715)	0.920	(0.867, 0.973)
500	0.724	(0.685, 0.763)	0.916	(0.892, 0.940)
1,000	0.689	(0.660, 0.718)	0.903	(0.885, 0.921)
5,000	0.707	(0.694, 0.720)	0.904	(0.896, 0.912)
10,000	0.703	(0.694, 0.712)	0.905	(0.899, 0.911)
50,000	0.709	(0.705, 0.713)	0.901	(0.898, 0.904)
100,000	0.707	(0.704, 0.710)	0.903	(0.901, 0.905)
150,000	0.706	(0.704, 0.708)	0.902	(0.900, 0.904)
Case 3				
r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.980	(0.930, 0.998)	1	(0.964, 1)
500	0.986	(0.976, 0.996)	0.998	(0.989, 1.000)
1,000	0.978	(0.969, 0.987)	0.995	(0.991, 0.999)
5,000	0.976	(0.972, 0.980)	0.993	(0.991, 0.995)
10,000	0.978	(0.975, 0.981)	0.995	(0.994, 0.996)
50,000	0.977	(0.976, 0.978)	0.995	(0.994, 0.996)
100,000	0.978	(0.977, 0.979)	0.995	(0.995, 0.995)
150,000	0.978	(0.977, 0.979)	0.995	(0.995, 0.995)

Table 4.8: NPI-RP-SO for the MW test with $H_1 : \mu_x \leq \mu_y \geq \mu_z$, $p = 2$, for Example 4.3

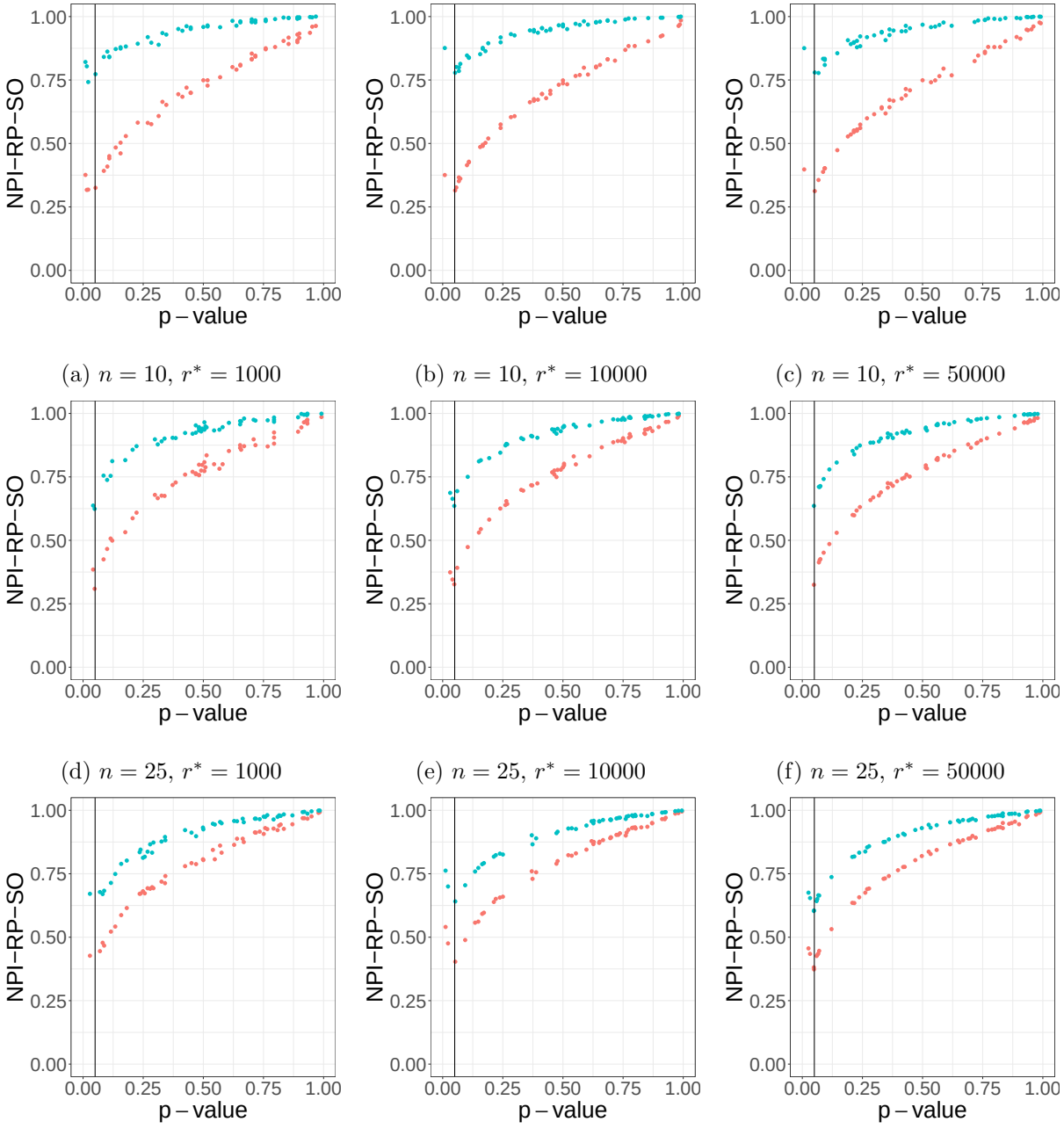
strong evidence in favour of the reproducibility of the original test result, hence, the results in Table 4.8 show high NPI reproducibility probabilities for Case 2 and Case 3. It can be concluded that reasonable approximations of the NPI lower and upper reproducibility probabilities can be obtained when the number of orderings sampled greater than or equal to 10,000.

4.7 Simulation studies

In this section, the reproducibility probability for the MW test and the EF test is explored via simulation. In Section 4.7.1, the reproducibility probability is investigated using the NPI-RP-SO approach. Section 4.7.2, considers the NPI-RP-B approach to study the reproducibility probability for the MW test and the EF test with known peak. In Section 4.7.3, the NPI-RP is investigated using NPI-RP-B for the MW test with unknown peak.

4.7.1 NPI-RP-SO simulation

This section investigates the reproducibility probability for the MW test and the EF test with $k = 3$ groups via simulation. The reproducibility is calculated using the NPI-RP-SO methodology, introduced in Section 4.4. The hypothesis of interests is $H_0 : \mu_x = \mu_y = \mu_z$ against $H_1 : \mu_x \leq \mu_y \geq \mu_z$, that is $p = 2$ which refers to the second group Y, the level of significance is $\alpha = 0.05$. The data were generated under H_0 and H_1. Under H_0, original data were generated from the Normal distribution with mean 0 and standard deviation 1. Under H_1, data were generated from Normal distribution with different means $\mu_x = 0$, $\mu_y = 1.5$ and $\mu_z = 1$, and standard deviation 1. To study the impact of the sample size on the lower and upper reproducibility probabilities, 50 replications of samples of sizes $n = 10, 25, 50$ were considered. Increasing the size of samples leads to increasing the power of the test, so we obtain more cases rejecting H_0 when simulations are performed by sampling under H_1. In Figures 4.1 and 4.2, the computation of the lower and upper reproducibility probabilities was achieved by sampling orderings of sizes $r^* = 1000, 10000, 50000$, to study the patterns of the NPI lower and upper reproducibility probabilities for different values of r^*. The vertical line indicates the test threshold value at $\alpha = 0.05$. Figures 4.1 and 4.2 show that there are no substantial differences on the patterns for different values of r^*. It is interesting to note that, when the sample size is small, the lower reproducibility probabilities seem to tend to 0.25 when the observed test statistics are close to the rejection region. Imprecision of reproducibility probabilities is equal to $\widehat{\overline{RP}}$ minus $\widehat{\underline{RP}}$. Figures 4.1 and 4.2 show that when the sample sizes increases the imprecision of reproducibility probabilities decreases.

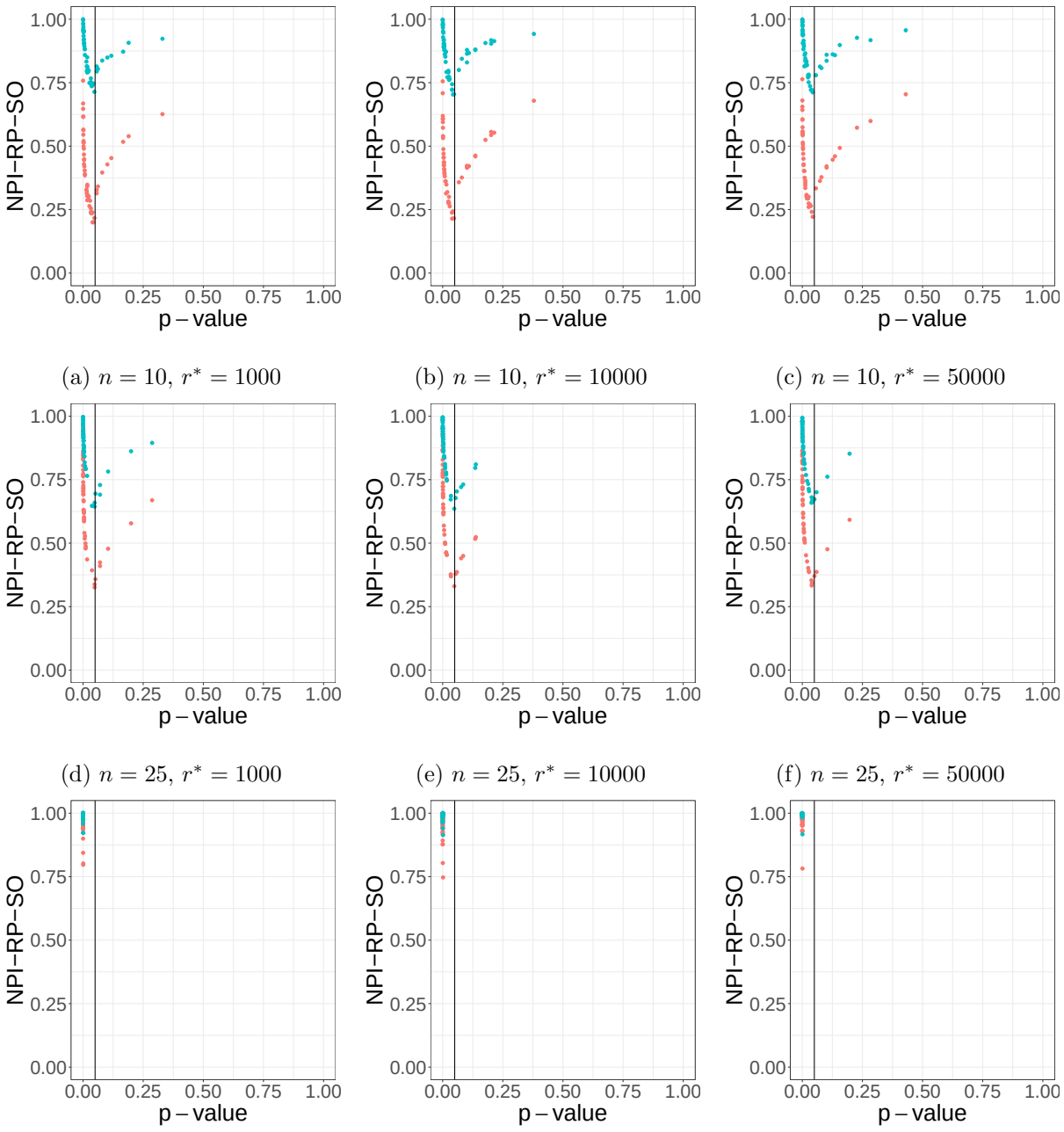


(a) $n = 10$, $r^* = 1000$ (b) $n = 10$, $r^* = 10000$ (c) $n = 10$, $r^* = 50000$

(d) $n = 25$, $r^* = 1000$ (e) $n = 25$, $r^* = 10000$ (f) $n = 25$, $r^* = 50000$

(g) $n = 50$, $r^* = 1000$ (h) $n = 50$, $r^* = 10000$ (i) $n = 50$, $r^* = 50000$

Figure 4.1: NPI-RP-SO under H_0, for simulated values of the upper (blue) and lower (red) RPs, for 50 replications, with $k = 3$ and the original samples from $N(0, 1)$, $\alpha = 0.05$



(a) $n = 10$, $r^* = 1000$ (b) $n = 10$, $r^* = 10000$ (c) $n = 10$, $r^* = 50000$

(d) $n = 25$, $r^* = 1000$ (e) $n = 25$, $r^* = 10000$ (f) $n = 25$, $r^* = 50000$

(g) $n = 50$, $r^* = 1000$ (h) $n = 50$, $r^* = 10000$ (i) $n = 50$, r^*=50000

Figure 4.2: NPI-RP-SO under H_1, for simulated values of the upper (blue) and lower (red) RPs, for 50 replications, with $k = 3$ and $X \sim N(0, 1)$, $Y \sim N(1.5, 1)$, $Z \sim N(1, 1)$, $\alpha = 0.05$

4.7.2 NPI-RP-B simulation with known peak

This section studies the reproducibility probability for the MW test and the EF test with known peak via simulations, where reproducibility is calculated using Algorithm 2. The NPI-RP-B method is preformed with infinite support (Approach II). The null hypothesis is H_0 : $\mu_1 = \mu_2 = ... = \mu_k$ and the alternative hypothesis is $H_1 : \mu_1 \leq \mu_2 \leq ... \leq \mu_{p-1} \leq \mu_p \geq \mu_{p+1} \geq$

Case	k	Simulation
1	3	$N(0,1)$
2	3	$X \sim N(0.6,1)$, $Y \sim N(1,1)$, $Z \sim N(0.5,1)$
3	3	$X \sim N(0.6,1)$, $Y \sim N(1.5,1)$, $Z \sim N(0.5,1)$
4	3	$X \sim N(0.6,1)$, $Y \sim N(2,1)$, $Z \sim N(0.5,1)$
5	3	Gamma(2, 1)
6	5	$N(0,1)$
7	5	$X \sim N(0.1,1)$, $Y \sim N(0.2,1)$, $Z \sim N(0.5,1)$, $V \sim N(0.2,1)$, $W \sim N(0.1,1)$

Table 4.9: Simulation cases for the MW test and the EF test

$... \geq \mu_{k-1} \geq \mu_k$. The level of significance is $\alpha = 0.05$. Data were simulated under H_0 and under H_1, as presented in Table 4.9. To study the impact of the number of groups and the sample size on the reproducibility probability, the simulation is considered with the number of groups $k = 3$, 5 and the sample sizes $n = 10$, 25. Each case introduced in Table 4.9 is considered with the sample sizes $n = 10$, 25. For $k = 3$ groups, $p = 2$ which refers to the second group. For $k = 5$ groups, $p = 3$ which refers to the third group. The reproducibility probability estimates are identical for both the MW test and the EF test with $k = 3$ groups and $p = 2$. However, with $k = 5$, the MW test statistic and the EF test statistic are differ because the Mann-Whitney sums used in the EF test statistic calculation are not uniformly weighted with a value of 1, as they are with $k = 3$. This variation in the test statistics results in varying reproducibility probability values for the two tests.

The inputs for the simulation study in Tables 4.10 through Table 4.23 are as follows: Algorithm 2 is applied with $B = 1000$ and $T = 100$. For each run, one sample of size n is generated from each of the distributions given in the Table 4.9, the MW test and the EF test are performed both on the these samples, and the tests outcomes are obtained and the RP estimates for the MW test and the EF test are calculated. The reproducibility probability estimates have been reported for 10 simulated original samples in each table. Note that the threshold values, introduced in Sections 4.2.1 and 4.2.2, are provided in the caption of each table for both tests. For the same test statistic, reproducibility probability estimates differs from one data set to another data set. These small variations in the RP estimates are due to variations in the original samples and in the NPI-B samples.

The relationship between NPI-RP-B and the p-value for the MW test and the EF test is examined in the simulations. We use the p-value for better visualization of figures rather than the critical value because each simulation scenario has a different critical value. This due to

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
141	1.804	0.036	R	0.480	0.512	0.512	0.548
135	1.540	0.062	NR	0.596	0.626	0.626	0.663
131	1.364	0.086	NR	0.597	0.627	0.629	0.660
125	1.100	0.136	NR	0.677	0.713	0.715	0.743
117	0.748	0.227	NR	0.733	0.763	0.763	0.795
108	0.352	0.362	NR	0.799	0.836	0.837	0.864
108	0.352	0.362	NR	0.804	0.833	0.833	0.858
91	-0.396	0.654	NR	0.903	0.920	0.920	0.939
77	-1.012	0.844	NR	0.960	0.972	0.972	0.985
59	-1.804	0.964	NR	0.985	0.992	0.992	0.999

Table 4.10: RP for the MW test and the EF test, with Case 1, $n = 10$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$

the variations in the sample sizes and the number of groups considered. Note that, the level of significance $\alpha = 0.05$ is represented on the figures by a vertical line. The observed p-value and the NPI-RP-B estimates for 100 data sets are displayed in Figures 4.3 through 4.9. From these figures, the NPI-RP-B estimates are low when the p-value is close to the threshold $\alpha = 0.05$ and it tend to be lower when the null hypothesis is rejected more than when it is not rejected. The reason for that is the presence of some sort of direction in the alternatives. Typically with p-value not close to the threshold, the data provide a strong evidence in favour of the reproducibility of the original test result. Similar findings have been observed in the previous NPI studies of test reproducibility [2, 18, 74, 97], and in Chapters 2 and 3. For the cases in Figures 4.4, 4.5 and 4.6, increasing the means for the peak group Y from $\mu_y = 1$, $\mu_y = 1.5$ to $\mu_y = 2$, leads to increase the power of the test and increase the number cases where the null hypothesis is rejected. Moreover, under H_1, increasing the size of samples from 10 to 25 leads to increasing the power of the test and more rejections of the null hypothesis.

Simulation studies show that there is an influence of sample size on the variability of NPI-RP estimates, meaning that NPI-RP-B values based on large samples sizes have less variability than NPI-RP estimates based on small sample sizes. Thus, with $n = 25$ the reproducibility probability curve becomes progressively smoother.

4.7.3 NPI-RP-B simulation with unknown peak

This section studies the NPI reproducibility probability for the MW test with unknown peak using the NPI-RP-B method with infinite support, as introduced in Section 4.5. In this section

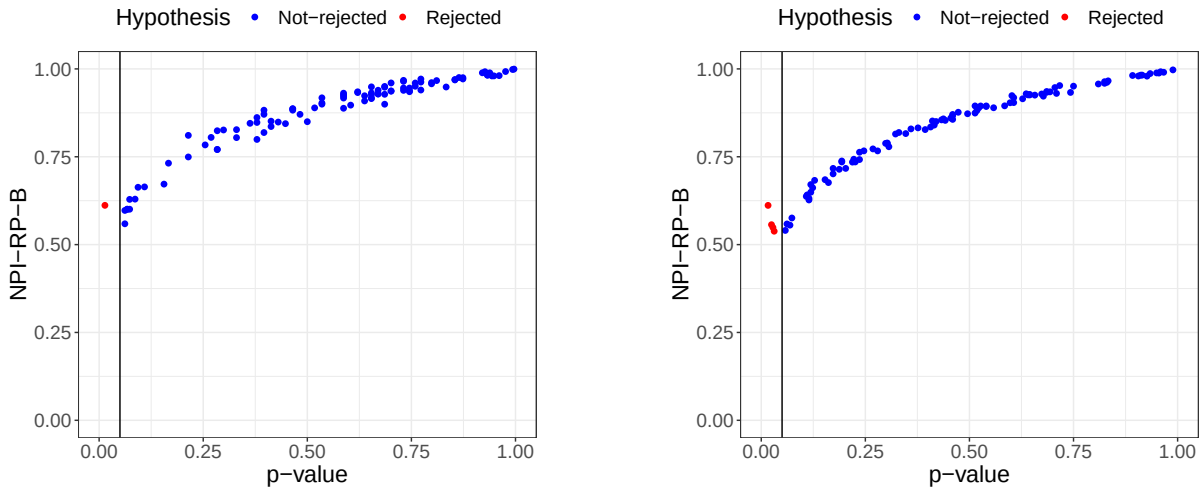
A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
773	1.663	0.048	R	0.425	0.478	0.478	0.521
744	1.337	0.091	NR	0.563	0.604	0.602	0.641
717	1.034	0.151	NR	0.632	0.673	0.673	0.712
693	0.764	0.222	NR	0.698	0.735	0.736	0.778
637	0.135	0.446	NR	0.823	0.861	0.861	0.888
609	-0.180	0.571	NR	0.884	0.909	0.909	0.935
609	-0.180	0.571	NR	0.852	0.880	0.879	0.909
555	-0.787	0.784	NR	0.946	0.959	0.959	0.972
473	-1.708	0.956	NR	0.979	0.989	0.988	0.998
395	-2.585	0.995	NR	0.987	0.996	0.997	1

Table 4.11: RP for the MW test and the EF test, with Case 1, $n = 25$, $\alpha = 0.05$, $A_{2,0.0499} = 772$, $Z_{0.05} = 1.645$

Algorithm 2 is applied with $B = 1000$ and $T = 1$. The NPI reproducibility probability is considered for $k = 3$ groups, X, Y and Z with $n = 10$. We test the hypothesis $H_0 : \mu_1 = \mu_2 = ... = \mu_k$ against $H_1 : \mu_1 \leq \mu_2 \leq ... \leq \mu_{p-1} \leq \mu_p \geq \mu_{p+1} \geq ... \geq \mu_k$, and when p is unknown. The level of significance is set at $\alpha = 0.05$, and it leads to the critical value $A_{0.0498} = 2.112$, which can be found in [20]. For the MW test with unknown peak and with large sample sizes the Monte Carlo Approximation (with 10000 iterations) is used to obtain the critical value $A_{0.0498} = 2.112$. The null hypothesis is rejected if $A'_{\hat{p}} \geq 2.112$.

In Table 4.24, the original data were generated from Normal distributions with different means $\mu_x = 0$, $\mu_y = 1.5$ and $\mu_z = 1$, and standard deviation 1. In Table 4.25, the original data were generated from Normal distributions with different means $\mu_x = 0.6$, $\mu_y = 1.5$ and $\mu_z = 0.5$, and the standard deviation 1. In Table 4.26, the original data were generated form Gamma distributions with different shape parameters $\theta_x = 1.5$, $\theta_y = 3$ and $\theta_z = 2$, and the scale parameter 1. Tables 4.24, 4.25 and 4.26 present the NPI-RP estimates for 10 original samples, for each sample we compute the probability of obtaining the same conclusion as the original test out of the $B = 1000$ and how each peak contributes to this probability value, denoted by (NPI-RP-B).

For the original samples 1, in Table 4.24, the original sample peak is estimated to be the second group, $\hat{p} = 2$, and the original test conclusion is the rejection of the null hypothesis, and the probability of rejection of the future tests in the 1000 bootstrapped samples is 0.992, and in this case the future samples with the peak is the second group contribute the most in this NPI-RP-B estimate, with value of 0.984, the future samples with the peak is the third



(a) $n = 10$ (b) $n = 25$

Figure 4.3: NPI-RP-B for the MW test and the EF test, with Case 1, $\alpha = 0.05$

group contribute with a value of 0.008, while there is no future samples with peak $\hat{p} = 1$. In addition, for the original samples 9, in Table 4.26, the original sample peak is estimated to be the second group $\hat{p} = 2$, the original test conclusion is non-rejection of the null hypothesis, and the probability of non-rejection of the future tests in the 1000 bootstrapped samples is 0.550, and in this case the future samples with the peak is the second group contributes the most in the estimate of NPI-RP-B with probability of non-rejection in the future tests of 0.253, while for the future samples with the peak estimated to be the first group out of the 1000 bootstrap samples contributes with probability of non-rejection value of 0.072, and the probability of non-rejection of the future samples with the peak estimated to be the third group is 0.225.

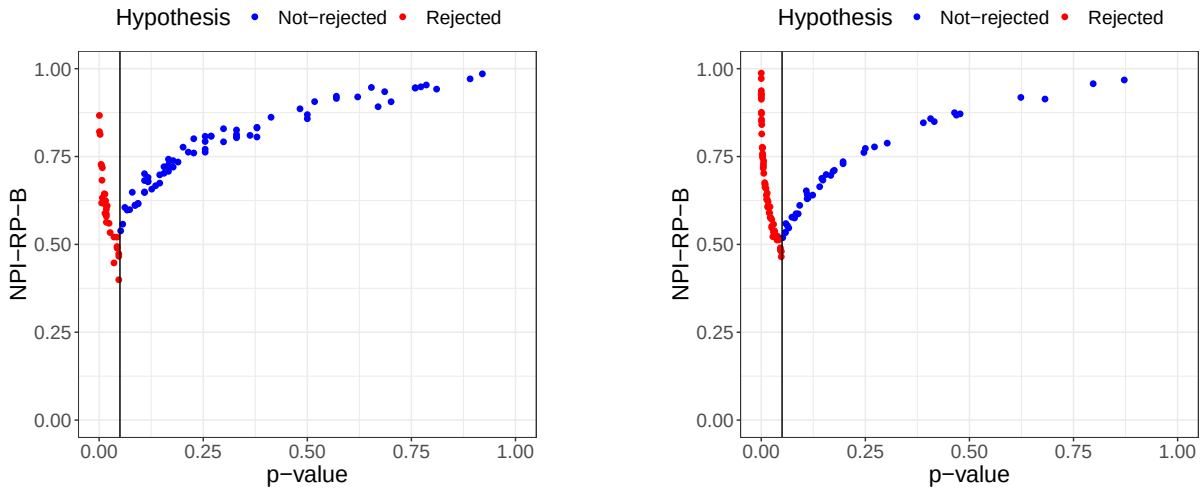
To sum up, the estimated peak group in the original sample, always contributes the most in the probability of obtaining the same conclusion as the original test in the future bootstrap samples. Again, the NPI reproducibility estimates are low when the when the test statistics are close to the threshold, and lower in cases of rejection than in non-rejection. The NPI-RP-B estimates are large when the original test statistics are away from the threshold.

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
152	2.288	0.011	R	0.577	0.615	0.615	0.669
145	1.980	0.024	R	0.514	0.549	0.549	0.589
139	1.716	0.043	R	0.385	0.417	0.415	0.450
136	1.584	0.057	NR	0.548	0.588	0.590	0.614
131	1.364	0.086	NR	0.582	0.619	0.619	0.650
129	1.276	0.101	NR	0.642	0.671	0.672	0.703
123	1.012	0.156	NR	0.707	0.737	0.737	0.772
119	0.836	0.202	NR	0.718	0.754	0.752	0.785
103	0.132	0.448	NR	0.857	0.886	0.887	0.915
88	-0.528	0.701	NR	0.921	0.942	0.943	0.964

Table 4.12: RP for the MW test and the EF test, with Case 2, $n = 10$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
907	3.169	0.001	R	0.806	0.841	0.842	0.872
846	2.484	0.006	R	0.642	0.692	0.693	0.719
808	2.057	0.020	R	0.558	0.598	0.598	0.643
786	1.810	0.035	R	0.484	0.523	0.525	0.560
768	1.607	0.054	NR	0.501	0.543	0.542	0.582
740	1.293	0.098	NR	0.570	0.614	0.614	0.662
736	1.248	0.106	NR	0.595	0.627	0.628	0.663
725	1.124	0.131	NR	0.614	0.663	0.664	0.702
667	0.472	0.318	NR	0.768	0.812	0.813	0.841
634	0.101	0.460	NR	0.829	0.861	0.859	0.890

Table 4.13: RP for the MW test and the EF test, with Case 2, $n = 25$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$



(a) $n = 10$ (b) $n = 25$

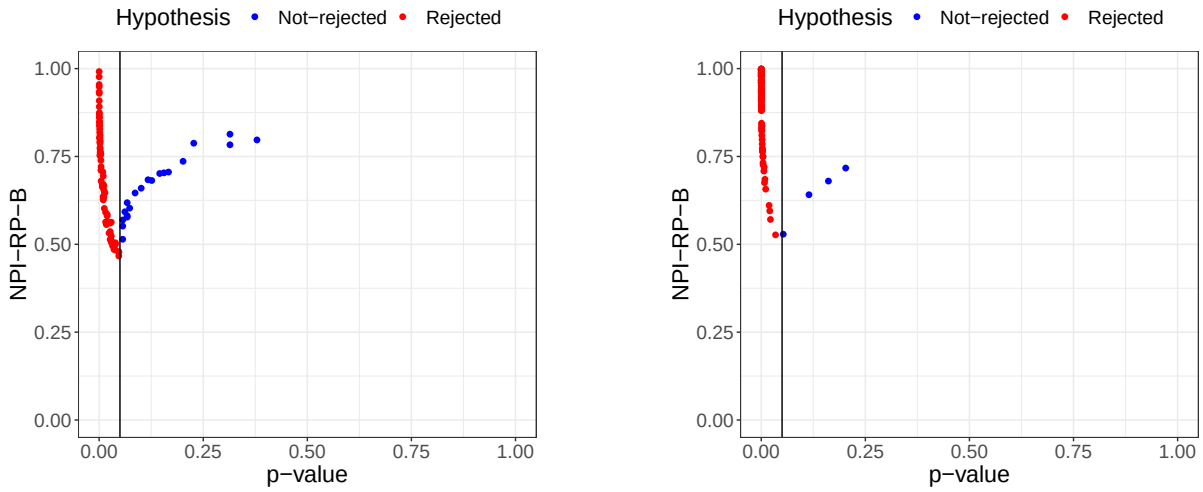
Figure 4.4: NPI-RP-B for the MW test and the EF test, with Case 2, $\alpha = 0.05$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
197	4.267	9.887×10^{-6}	R	0.988	0.995	0.995	0.999
185	3.740	9.219×10^{-5}	R	0.960	0.974	0.974	0.985
169	3.036	0.001	R	0.754	0.784	0.785	0.823
154	2.376	0.009	R	0.617	0.656	0.656	0.689
147	2.068	0.019	R	0.556	0.585	0.585	0.622
138	1.672	0.047	R	0.452	0.483	0.481	0.525
130	1.320	0.093	NR	0.604	0.638	0.635	0.672
130	1.320	0.093	NR	0.602	0.651	0.651	0.697
128	1.232	0.109	NR	0.667	0.698	0.698	0.731
112	0.528	0.299	NR	0.790	0.820	0.819	0.853

Table 4.14: RP for the MW test and the EF test, with Case 3, $n = 10$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
1057	4.855	6.011×10^{-7}	R	0.986	0.993	0.994	1
1032	4.574	2.389×10^{-6}	R	0.972	0.982	0.983	0.993
985	4.046	2.604×10^{-5}	R	0.940	0.957	0.958	0.969
940	3.540	1.998×10^{-4}	R	0.886	0.915	0.916	0.934
932	3.450	2.799×10^{-4}	R	0.874	0.899	0.899	0.915
902	3.113	9.253×10^{-4}	R	0.815	0.851	0.851	0.873
893	3.012	0.001	R	0.792	0.829	0.828	0.862
883	2.900	0.002	R	0.771	0.806	0.807	0.840
855	2.585	0.005	R	0.682	0.725	0.726	0.753
817	2.158	0.015	R	0.581	0.624	0.627	0.663

Table 4.15: RP for the MW test and the EF test, with Case 3, $n = 25$, $\alpha = 0.05$, $A_{2,0.0499} = 772$, $Z_{0.05} = 1.645$



(a) $n = 10$ (b) $n = 25$

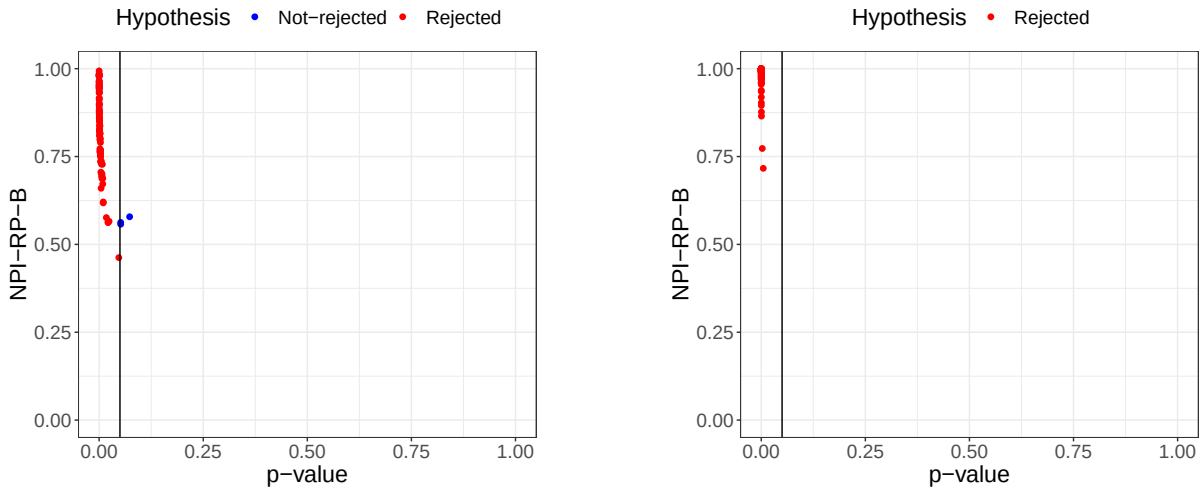
Figure 4.5: NPI-RP-B for the MW test and the EF test, with Case 3, $\alpha = 0.05$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
200	4.399	5.427×10^{-6}	R	0.995	0.999	0.999	1
180	3.520	2.162×10^{-4}	R	0.859	0.886	0.887	0.912
180	3.520	2.162×10^{-4}	R	0.916	0.941	0.941	0.956
174	3.256	5.658×10^{-4}	R	0.894	0.923	0.923	0.943
168	2.992	0.001	R	0.843	0.877	0.879	0.906
161	2.684	0.004	R	0.689	0.718	0.719	0.751
159	2.596	0.005	R	0.674	0.711	0.710	0.746
152	2.288	0.011	R	0.599	0.637	0.638	0.670
148	2.112	0.017	R	0.541	0.574	0.574	0.611
143	1.892	0.029	R	0.463	0.501	0.501	0.540

Table 4.16: RP for the MW test and the EF test, with Case 4, $n = 10$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
1145	5.844	2.544×10^{-9}	R	0.996	0.999	0.999	1
1125	5.620	9.575×10^{-9}	R	0.995	0.999	0.999	1
1102	5.361	4.138×10^{-8}	R	0.995	0.998	0.998	1
1087	5.192	1.038×10^{-7}	R	0.993	0.997	0.997	1
1071	5.013	2.685×10^{-7}	R	0.991	0.996	0.997	1
1045	4.720	1.177×10^{-6}	R	0.981	0.990	0.990	0.996
1020	4.439	4.510×10^{-6}	R	0.979	0.989	0.989	0.996
992	4.125	1.856×10^{-5}	R	0.954	0.967	0.968	0.979
971	3.889	5.039×10^{-5}	R	0.921	0.943	0.944	0.960
951	3.664	1.242×10^{-4}	R	0.901	0.926	0.927	0.944

Table 4.17: RP for the MW test and the EF test, with Case 4, $n = 25$, $\alpha = 0.05$, $A_{2,0.0499} = 772$, $Z_{0.05} = 1.645$



(a) $n = 10$ (b) $n = 25$

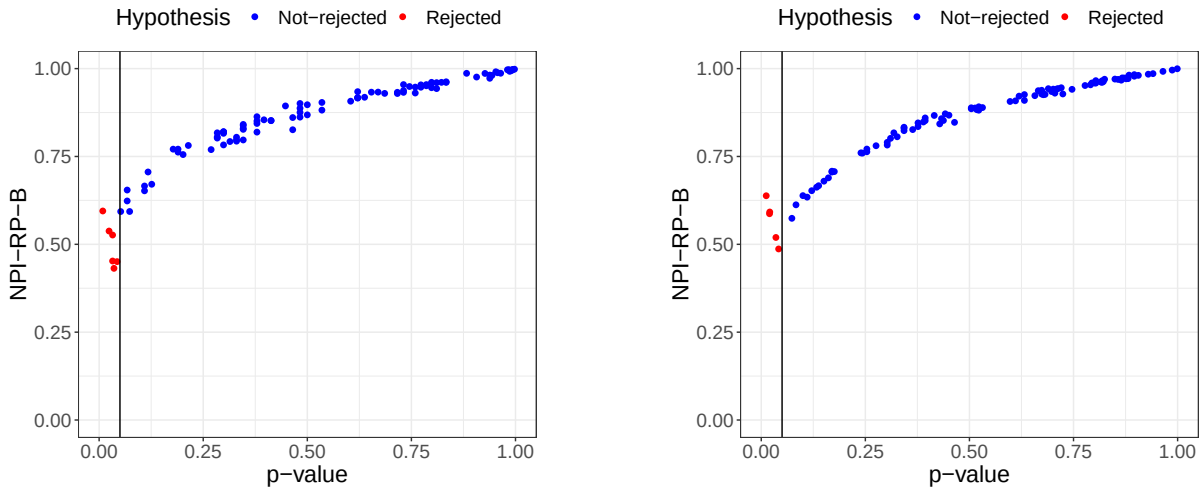
Figure 4.6: NPI-RP-B for the MW test and the EF test, with Case 4, $\alpha = 0.05$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
137	-1.056	0.855	NR	0.518	0.561	0.562	0.590
125	1.100	0.136	NR	0.690	0.722	0.722	0.753
121	0.924	0.178	NR	0.698	0.735	0.738	0.764
115	0.660	0.255	NR	0.749	0.783	0.786	0.813
103	0.132	0.448	NR	0.842	0.875	0.875	0.902
98	-0.088	0.535	NR	0.842	0.876	0.877	0.902
90	-0.440	0.670	NR	0.916	0.935	0.937	0.953
84	-0.704	0.759	NR	0.930	0.948	0.949	0.962
76	-1.056	0.855	NR	0.955	0.968	0.969	0.983
62	-1.672	0.953	NR	0.984	0.993	0.993	0.998

Table 4.18: RP for the MW test and the EF test, with Case 5, $n = 10$, $\alpha = 0.05$, $A_{2,0.0498} = 138$, $Z_{0.05} = 1.645$

A_p	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
735	1.236	0.108	NR	0.609	0.647	0.644	0.689
729	1.169	0.121	NR	0.621	0.658	0.657	0.699
710	0.955	0.170	NR	0.673	0.706	0.707	0.741
692	0.753	0.226	NR	0.701	0.738	0.741	0.774
664	0.438	0.331	NR	0.797	0.824	0.824	0.849
632	0.079	0.469	NR	0.840	0.862	0.863	0.890
597	-0.315	0.624	NR	0.877	0.900	0.899	0.920
555	-0.787	0.784	NR	0.924	0.948	0.949	0.962
504	-1.360	0.913	NR	0.965	0.984	0.985	0.993
467	-1.776	0.962	NR	0.986	0.994	0.994	0.999

Table 4.19: RP for the MW test and the EF test, with Case 5, $n = 25$, $\alpha = 0.05$, $A_{2,0.0499} = 772$, $Z_{0.05} = 1.645$



(a) $n = 10$ (b) $n = 25$

Figure 4.7: NPI-RP-B for the MW test and the EF test, with Case 5, $\alpha = 0.05$

A_p	p-value	H_0	Min	Mean	Median	Max	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
377	0.045	R	0.400	0.436	0.435	0.476	1.666	0.048	R	0.404	0.438	0.437	0.483
364	0.079	NR	0.580	0.616	0.617	0.650	1.283	0.100	NR	0.601	0.635	0.635	0.674
352	0.125	NR	0.670	0.711	0.714	0.743	1.180	0.119	NR	0.648	0.692	0.694	0.725
339	0.195	NR	0.728	0.769	0.770	0.804	0.826	0.204	NR	0.740	0.772	0.773	0.803
327	0.276	NR	0.755	0.785	0.787	0.808	0.590	0.278	NR	0.750	0.780	0.781	0.811
294	0.553	NR	0.868	0.889	0.889	0.913	-0.088	0.535	NR	0.864	0.887	0.887	0.911
294	0.553	NR	0.875	0.900	0.901	0.925	-0.206	0.582	NR	0.879	0.906	0.906	0.927
261	0.806	NR	0.943	0.955	0.956	0.974	-0.885	0.812	NR	0.946	0.957	0.958	0.976
207	0.980	NR	0.988	0.995	0.995	0.999	-2.049	0.980	NR	0.988	0.995	0.995	0.999
195	0.990	NR	0.992	0.998	0.998	1	-2.359	0.991	NR	0.992	0.998	0.998	1

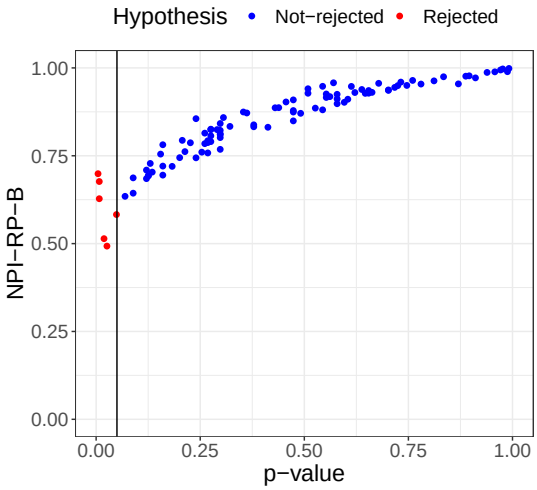
Table 4.20: RP for the MW test and the EF test, with Case 6, $n = 10$, $\alpha = 0.05$, $A_{2,0.0478} = 376$, $Z_{0.05} = 1.645$

A_p	p-value	H_0	Min	Mean	Median	Max	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
2177	0.045	R	0.450	0.490	0.490	0.525	1.742	0.041	R	0.455	0.500	0.502	0.546
2139	0.069	NR	0.537	0.572	0.572	0.613	1.502	0.067	NR	0.538	0.569	0.569	0.615
2126	0.079	NR	0.561	0.589	0.589	0.623	1.430	0.076	NR	0.554	0.586	0.587	0.621
2106	0.097	NR	0.597	0.630	0.631	0.661	1.284	0.100	NR	0.601	0.631	0.632	0.664
2096	0.107	NR	0.591	0.627	0.628	0.659	1.224	0.110	NR	0.600	0.630	0.630	0.664
2056	0.154	NR	0.671	0.705	0.706	0.742	1.029	0.152	NR	0.668	0.702	0.703	0.737
1936	0.366	NR	0.813	0.839	0.840	0.860	0.349	0.363	NR	0.800	0.834	0.835	0.855
1821	0.619	NR	0.872	0.894	0.894	0.915	-0.282	0.611	NR	0.873	0.894	0.895	0.916
1603	0.937	NR	0.980	0.988	0.989	0.996	-1.513	0.935	NR	0.979	0.987	0.987	0.995
1484	0.986	NR	0.993	0.998	0.998	1	-2.151	0.984	NR	0.992	0.997	0.997	1

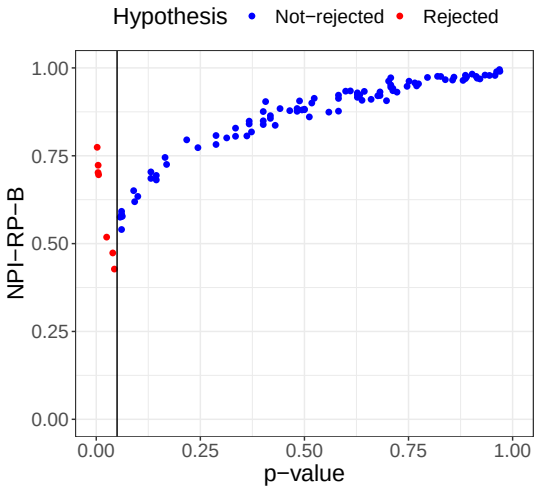
Table 4.21: RP for the MW test and the EF test, with Case 6, $n = 25$, $\alpha = 0.05$, $A_{2,0.0499} = 2168$, $Z_{0.05} = 1.645$

A_p	p-value	H_0	Min	Mean	Median	Max	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
448	0.001	R	0.814	0.842	0.841	0.868	3.199	0.001	R	0.814	0.844	0.844	0.872
409	0.008	R	0.597	0.625	0.625	0.673	2.330	0.010	R	0.582	0.619	0.620	0.668
380	0.039	R	0.467	0.497	0.495	0.532	1.784	0.037	R	0.486	0.516	0.516	0.555
376	0.047	R	0.446	0.485	0.486	0.522	1.651	0.049	R	0.449	0.484	0.483	0.518
370	0.061	NR	0.538	0.585	0.587	0.614	1.578	0.057	NR	0.536	0.576	0.577	0.603
361	0.089	NR	0.614	0.652	0.653	0.690	1.312	0.095	NR	0.608	0.650	0.651	0.687
347	0.150	NR	0.663	0.695	0.696	0.730	1.076	0.141	NR	0.655	0.689	0.689	0.724
334	0.226	NR	0.747	0.774	0.775	0.807	0.752	0.226	NR	0.740	0.765	0.765	0.799
306	0.447	NR	0.829	0.862	0.863	0.891	0.074	0.471	NR	0.830	0.865	0.865	0.896
260	0.812	NR	0.926	0.949	0.950	0.965	-0.840	0.800	NR	0.922	0.947	0.948	0.965

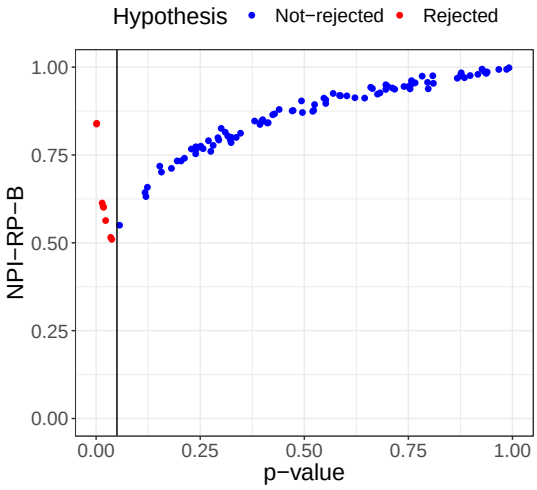
Table 4.22: RP for the MW test and EF test, with Case 7, $n = 10$, $\alpha = 0.05$, $A_{2,0.0478} = 376$, $Z_{0.05} = 1.645$



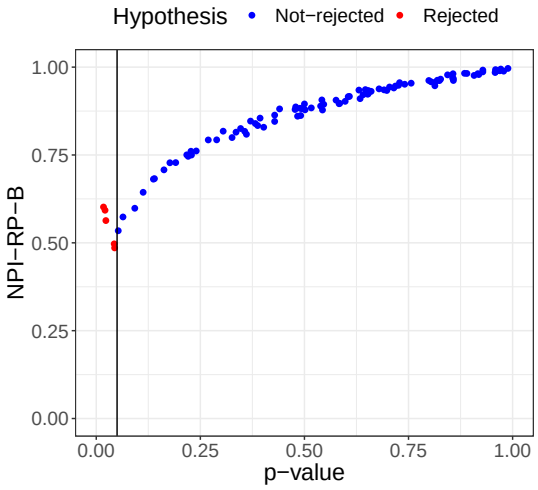
(a) MW, $n = 10$



(b) EF, $n = 10$



(c) MW, $n = 25$

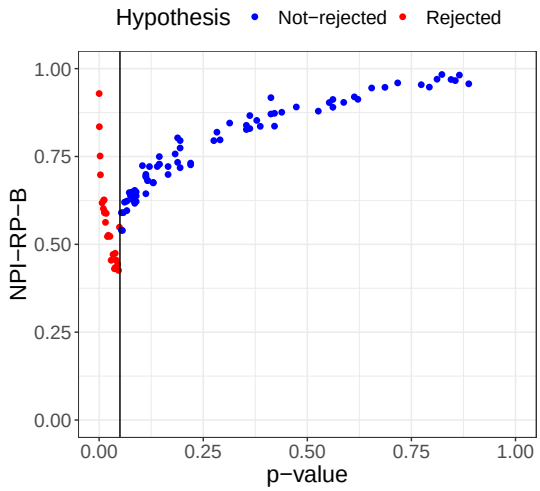


(d) EF, $n = 25$

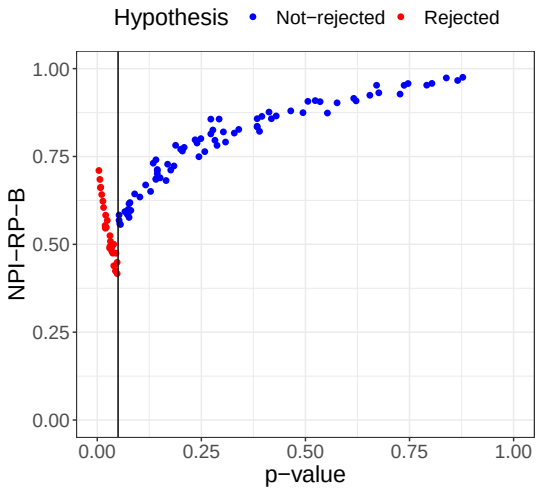
Figure 4.8: NPI-RP-B under H_0, with Case 6, $\alpha = 0.05$

A_p	p-value	H_0	Min	Mean	Median	Max	$\tilde{A}_p^*$	p-value	H_0	Min	Mean	Median	Max
2483	3.105×10^{-4}	R	0.842	0.871	0.873	0.892	3.334	4.279×10^{-4}	R	0.834	0.862	0.864	0.884
2382	0.002	R	0.740	0.765	0.764	0.800	2.808	0.002	R	0.735	0.761	0.760	0.793
2341	0.004	R	0.702	0.733	0.733	0.755	2.681	0.004	R	0.722	0.744	0.744	0.771
2313	0.007	R	0.683	0.707	0.705	0.735	2.478	0.007	R	0.687	0.712	0.712	0.738
2250	0.017	R	0.555	0.595	0.594	0.632	2.099	0.018	R	0.553	0.594	0.592	0.634
2196	0.035	R	0.466	0.520	0.520	0.558	1.851	0.032	R	0.481	0.532	0.533	0.570
2118	0.086	NR	0.547	0.583	0.583	0.619	1.382	0.084	NR	0.546	0.579	0.580	0.616
2020	0.207	NR	0.718	0.756	0.758	0.781	0.788	0.215	NR	0.723	0.754	0.756	0.781
1838	0.583	NR	0.896	0.917	0.917	0.936	-0.203	0.580	NR	0.892	0.911	0.910	0.931
1790	0.684	NR	0.927	0.945	0.946	0.960	-0.454	0.675	NR	0.925	0.942	0.942	0.960

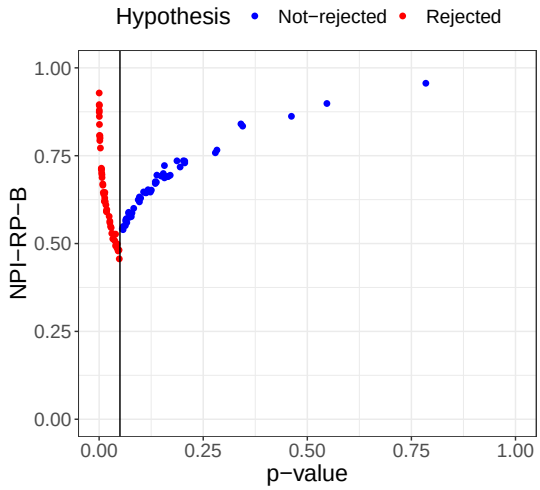
Table 4.23: RP for the MW test and the EF test, with Case 7, $n = 25$, $\alpha = 0.05$, $A_{2,0.0499} = 2168$, $Z_{0.05} = 1.645$



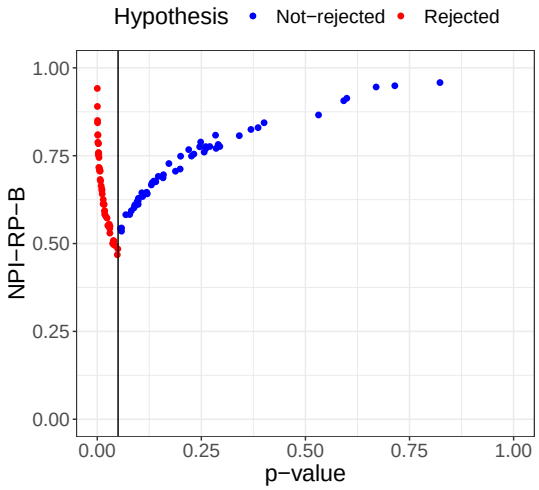
(a) MW, $n = 10$



(b) EF, $n = 10$



(c) MW, $n = 25$



(d) EF, $n = 25$

Figure 4.9: NPI-RP-B under H_1, with Case 7, $\alpha = 0.05$

	Test conclusion			Rejection				Non-rejection			
Samples	$\hat{p}$	$A'_{\hat{p}}$	H_0	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
1	2	4.267	R	0.992	0.000	0.984	0.008	0.008	0.000	0.006	0.002
2	3	3.384	R	0.812	0.000	0.041	0.771	0.188	0.007	0.021	0.160
3	3	3.004	R	0.867	0.000	0.257	0.610	0.133	0.008	0.053	0.072
4	2	2.948	R	0.746	0.001	0.618	0.127	0.254	0.012	0.133	0.109
5	2	2.772	R	0.855	0.002	0.587	0.266	0.145	0.008	0.092	0.045
6	2	2.640	R	0.782	0.004	0.561	0.217	0.218	0.013	0.152	0.053
7	2	2.552	R	0.888	0.001	0.436	0.451	0.112	0.002	0.051	0.059
8	2	2.376	R	0.560	0.013	0.477	0.070	0.440	0.059	0.261	0.120
9	3	2.281	R	0.587	0.001	0.113	0.473	0.413	0.023	0.098	0.292
10	2	1.628	NR	0.369	0.017	0.278	0.074	0.631	0.124	0.306	0.201

Table 4.24: NPI-RP-B for the MW test with unknown peak, $k = 3$, $X \sim N(0,1)$, $Y \sim N(1.5,1)$, $Z \sim N(1,1)$, $n = 10$, $\alpha = 0.05$, $A_{0.0498} = 2.112$, $B = 1000$, $T = 1$

	Test conclusion			Rejection				Non-rejection			
Samples	$\hat{p}$	$A'_{\hat{p}}$	H_0	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
1	2	3.696	R	0.966	0.074	0.890	0.002	0.034	0.009	0.025	0.000
2	2	3.564	R	0.894	0.017	0.856	0.021	0.106	0.014	0.079	0.013
3	2	3.256	R	0.817	0.046	0.763	0.008	0.183	0.024	0.142	0.017
4	2	2.904	R	0.741	0.056	0.666	0.019	0.259	0.050	0.183	0.026
5	2	2.860	R	0.691	0.023	0.646	0.022	0.309	0.050	0.210	0.049
6	2	2.464	R	0.630	0.182	0.440	0.008	0.370	0.113	0.232	0.025
7	2	2.464	R	0.625	0.086	0.525	0.014	0.375	0.098	0.243	0.034
8	2	2.244	R	0.506	0.050	0.442	0.014	0.494	0.135	0.292	0.067
9	2	2.200	R	0.486	0.031	0.427	0.028	0.514	0.104	0.313	0.097
10	2	1.672	NR	0.397	0.033	0.291	0.073	0.603	0.128	0.308	0.167

Table 4.25: NPI-RP-B for the MW test with unknown peak, $k = 3$, $X \sim N(0.6,1)$, $Y \sim N(1.5,1)$, $Z \sim N(0.5,1)$, $n = 10$, $\alpha = 0.05$, $A_{0.0498} = 2.112$, $B = 1000$, $T = 1$

	Test conclusion			Rejection				Non-rejection			
Samples	$\hat{p}$	$A'_{\hat{p}}$	H_0	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$	NPI-RP-B	$\hat{p}=1$	$\hat{p}=2$	$\hat{p}=3$
1	2	4.004	R	0.910	0.005	0.885	0.020	0.090	0.008	0.065	0.017
2	2	3.476	R	0.858	0.001	0.754	0.103	0.142	0.003	0.099	0.040
3	2	3.476	R	0.837	0.001	0.739	0.097	0.163	0.012	0.110	0.041
4	2	3.036	R	0.694	0.011	0.639	0.044	0.306	0.032	0.215	0.059
5	2	2.948	R	0.699	0.024	0.641	0.034	0.301	0.047	0.206	0.048
6	2	2.508	R	0.611	0.014	0.494	0.103	0.389	0.038	0.241	0.110
7	2	2.376	R	0.594	0.007	0.483	0.104	0.406	0.038	0.239	0.129
8	2	2.244	R	0.533	0.088	0.424	0.021	0.467	0.122	0.286	0.059
9	2	1.672	NR	0.450	0.008	0.284	0.158	0.550	0.072	0.253	0.225
10	2	1.452	NR	0.371	0.013	0.241	0.117	0.629	0.093	0.270	0.266

Table 4.26: NPI-RP-B for the MW test with unknown peak, $k = 3$, $X \sim$ Gamma(1.5, 1), $Y \sim$ Gamma(3, 1), $Z \sim$ Gamma(2, 1), $n = 10$, $\alpha = 0.05$, $A_{0.0498} = 2.112$, $B = 1000$, $T = 1$

4.8 Concluding remarks

This chapter explored the NPI reproducibility for the Mack-Wolfe (MW) test and the Esra-Fikri (EF) test. The exact NPI lower and upper reproducibility probabilities for the MW test are derived for three groups; however, for more than three groups and large samples, going through all possible orderings is computationally expensive. To this end, two NPI-based approaches are implemented, namely, the sampling of orderings (NPI-RP-SO) and the NPI-bootstrap (NPI-RP-B).

The NPI-RP-SO and NPI-RP-B methods has been applied to a variety of scenarios via simulation studies. The investigation in this chapter implies that the NPI reproducibility is quite poor, if the test statistic is close to the threshold, particularly, when H_0 is rejected more than when H_0 is not rejected. The reason for that is the presence of some sort of direction in the alternatives. This pattern was observed in our investigation of NPI-RP for ordered alternatives, as detailed in Chapter 3.

Further, the NPI-RP results for the MW test and EF test with three groups are identical using the NPI-RP-E, NPI-RP-SO and NPI-RP-B methods, as the EF test statistic with three groups is the sum of two Mann-Whitney statistics, each with weight 1.

This chapter contributes to the development of NPI reproducibility which was introduced by Coolen and BinHimd [32], and the findings of the NPI reproducibility probability for the MW test and the EF test are consistent with previous NPI studies of test reproducibility

[2, 18, 33, 97], where the reproducibility is low close the the boundaries of the rejection region.

There are many research challenges for the further development of NPI for reproducibility of tests. For example, in this chapter the reproducibility of the MW test for more than three groups was investigated using the NPI-B approach. However, generalizing the MW test's exact lower and upper reproducibility probabilities for more than three groups is of interest for future work which may require developing some methods. The reproducibility for other umbrella alternatives tests such as the Chen and Wolfe [22] test and Hettmansperger and Norton [56] test, can also be studied. The parametric predictive bootstrap method, introduced by Aldawsari [2], can also be used to study the reproducibility and possibly compared to the results in this chapter.

Chapter 5

Reproducibility of Slippage Tests

5.1 Introduction

In the previous chapters, we introduced NPI reproducibility for several settings including the general alternatives tests, ordered alternatives tests and umbrella alternatives tests. This chapter contributes to the development of NPI reproducibility for statistical hypothesis tests by considering the reproducibility of one of the slippage tests, namely, the Mosteller test. Slippage tests are designed to diagnose whether one or more groups are slipped either to the right or to the left relative to the remaining groups. Mosteller test is used to test the null hypothesis that all groups are identical against the alternative hypothesis that one of the groups has slipped to the right or to the left of the rest.

Section 5.2 presents an overview of the slippage tests. Section 5.3 provides a review of the Mosteller test. Section 5.4 introduces the concept of strong reproducibility for the Mosteller test. In Section 5.5, strong reproducibility is considered using the NPI approach to derive the exact NPI lower and upper reproducibility probabilities for the Mosteller test. This is achieved by considering all orderings of future observations among the data observations from each group, which are equally likely. However, it is computationally challenging to derive such lower and upper probabilities for data sets with large sample size and number of groups, due to the increase in the number of orderings of future observations among the data observations, resulting in an increase in the computational time. Section 5.6 presents the NPI sampling of orderings approach to investigate strong reproducibility for the Mosteller test. Section 5.7 provides the NPI reproducibility probability for the Mosteller test using the NPI-RP-B approach. In section 5.8, application examples are provided. The NPI reproducibility probability for the Mosteller test is investigated via simulation in section 5.9. We conclude the content of this chapter in Section 5.10.

5.2 Slippage tests

Slippage tests are designed to test the null hypothesis that the means are the same against the alternative hypothesis that one or more means are slipped from the others, either to the right or to the left. The term 'slipped to the right' is used when the observations of one or multiple groups tend to be larger than the observations in the other groups. The term 'slipped to the left' is used when the observations of one or multiple groups tend to be smaller than the observations in the other groups. The term 'a general test for slippage' is used to describe a test where an unspecified subset of groups have slipped. A more general problem would be when the direction of the slippage is not specified [52]. A 'specific test for slippage' is one where only a number of the groups have slipped.

Mosteller [79] proposed a specific slippage test for the case of k groups of equal size n, to determine whether one of k groups has slipped to the right or to the left of the rest, and this test has become known in the literature as Mosteller's k-sample slippage test. Under the null hypothesis, all groups are continuous and identical [79]. Bofinger [19] investigated some properties of the Mosteller test and introduced a generalization of this test to the problem of whether a subset of the k groups has slipped. If slippage to the left is of interest, the test procedure is to count the number of observations in the subset groups possessing the smallest minimum which are smaller than all observations in the remaining groups. Mosteller and Tukey [80] proposed a method to calculate the slippage test when the samples are of unequal size.

Another test of this type has been suggested by Granger and Neave [52] for the k-sample slippage problem with one group had slipped. This test does not require special tables and have satisfactory power. Neave [82, 83] presented four simple tests for slippage in k-sample situations, with clear distinctions being made between different forms of the alternative hypothesis. Those tests were designed for the alternative hypothesis that precisely one group had slipped. The four alternative hypotheses are: slippage of a specified group in a specified direction, slippage of a specified group in either direction, slippage of any group in a specified direction and slippage of any group in either direction.

Doornbos and Prins [42] discussed slippage tests for a variety of distributions, such as the Normal, the Poisson, the Binomial and the Negative Binomial, as well as a distribution free k-sample Slippage test. A similar problem dealing with the means of several Normal distributions has been investigated by Paulson [88].

Slippage problems have been considered in the literature for the variances of k groups. Traux [105] introduced optimum procedure which is subject to certain restrictions to decide whether all groups variances are equal, and if not, which has the largest or smallest variance. This

procedure maximises the probability of making the correct decision. Doornbos and Prins [41] introduced slippage tests for a set of estimated normal variances. Further, Doornbos and Prins [41] considered the power function of these tests with respect to the alternative hypothesis that one of the variances has slipped to the right or to the left.

In this chapter, we consider the reproducibility probability for the Slippage test for the location problem, proposed by Mosteller [79], as it is the most well-known test for the slippage problem. Mosteller [79] considers the specific alternative that one group has slipped. We will focus on the case when the sample sizes are equal. However, the reproducibility probability for the Mosteller test with unequal sample sizes is also consider for some scenarios.

5.3 Mosteller test

Mosteller [79] proposed a slippage test which is designed to test the null hypothesis that all k groups are identical against the alternative that one group has slipped either to the right or to the left. The term slipped to the right is used when a given group has the largest observations than the other groups. The term slipped to the left is used when a given group has the smallest observations than the other groups. In this chapter, we will focus of the alternative hypothesis that one group has slipped to the left.

If we are interested in detecting a slippage to the left of any one group, the testing process involves sorting all the observations in the groups from smallest to largest. Then, select the group with the smallest observation and count the number of observations, r, in this group that are smaller than all the observations in the remaining $k-1$ groups. Let $\mathcal{R}$ be a random variable representing the number of observations from a group (the group that provides the smallest observations among k groups) that are smaller than all observations from the remaining $k-1$ groups [19, 79]. When the samples are of equal size , $P(\mathcal{R} \geq r)$, is given by

$$P_{(r)} = P(\mathcal{R} \geq r) = \frac{k(n!)(kn-r)!}{(kn)!(n-r)!} \tag{5.1}$$

For very large n

$$P_{(r)} \approx \frac{1}{k^{1-r}} \tag{5.2}$$

The null hypothesis is rejected at level of significance α, if

$$P_{(r)} \leq \alpha$$

The Mosteller test has the advantage of being easy to implement, however, it requires

samples of equal sizes [79]. When the samples are of unequal size [80], then

$$P_{(r)} = \frac{\sum_i n_i^{(r)}}{N^{(r)}} \tag{5.3}$$

where n_i is the number of observations in sample i, $i = 1, ..., k$, and $N = \sum_{i=1}^{k} n_i$. Consider the r smallest values, $n_i^{(r)} = n_i(n_i - 1)\dots(n_i - r + 1)$ and $N^{(r)} = N(N-1)\dots(N-r+1)$ [80].

There is symmetry between slippage to the right and slippage to the left. Thus, if we are interested in detecting a slippage to the right of any one group, the same methods in Equations (5.1) and (5.3) can be used to find the probability that the group with the largest observation has r or more observations which preceded all observations in the other $k-1$ groups. The testing process involves sorting all the observations in the groups from largest to smallest. Then, select the group with the largest observation and count the number of observations, r, in this group that are larger than all the observations in the remaining $k-1$ groups.

In classical tests there are two types of errors that are not sufficient to illustrate the situation with the Mosteller test, these errors are: rejecting the null hypothesis when it is true and accepting the null hypothesis when it is false. There is a third type of error because the Mosteller test depends on the idea of making the correct decision about which group has slipped to the right or to the left [79]. We may make the error of correctly rejecting the null hypothesis based on the wrong reason. This means it is possible for the null hypothesis to be false. It is also possible to reject the null hypothesis when a group has too many observations that are smaller or greater than all observations in the other groups, but the group from which another group is drawn is in fact left or right most group. In this situation, the third type of error is committed [79].

5.4 Strong reproducibility probability

The reproducibility probability is the probability of the event that, if a statistical test were repeated, under the same circumstances, the same conclusion as the original test would reached, with regard to rejection or non-rejection of the null hypothesis. The concept of strong reproducibility is considered within the context of the Mosteller test, particularly when the null hypothesis is rejected, suggesting that one group has slipped. Strong reproducibility means reproducing the rejection of the null hypothesis with the same group that slipped in the original data is also slipped for the future data. In Sections 5.5, 5.6 and 5.9, strong reproducibility will be considered using the NPI-RP-E, NPI-RP-SO and NPI-RP-B approaches.

5.5 NPI-RP-E for the Mosteller test

In this section, NPI reproducibility probability for the Mosteller test is introduced in term of the lower and upper reproducibility, denoted by $\underline{RP}$ and $\overline{RP}$, respectively. Suppose there are $k \geq 2$ independent groups, and the Mosteller test is performed. We assume that the null hypothesis is rejected in the original test with $P_{(r)} \leq \alpha$, that is, group l^* has slipped to the left, so r is the number of observations from group l^* that are smaller than the smallest observation from all other groups l, where $l \neq l^*$. We will consider strong reproducibility, that the null hypothesis is rejected in the future test with the the same group has slipped to the left.

Let $x_1^l < x_2^l < \ldots < x_{n_l}^l$ be the ordered observed values of group l, $l = 1, 2, ..., k$ where $l \neq l^*$. These observations partition the real-line into $n_{n_l}+1$ intervals, $I_{i_l}^l = (x_{i_l-1}^l, x_{i_l}^l)$, $i_l = 1, 2, \ldots, n_l + 1$. Let $x_1^{l^*} < x_2^{l^*} < \ldots < x_{n_{l^*}}^{l^*}$ be the ordered observed values of group l^*, these observations partition the real-line into $n_{n_{l^*}} + 1$ intervals, $I_{i_{l^*}}^{l^*} = (x_{i_{l^*}-1}^{l^*}, x_{i_{l^*}}^{l^*})$, $i_{l^*} = 1, 2, \ldots, n_{l^*} + 1$. For ease of notation, let $x_0^l = x_0^{l^*} = -\infty$ and $x_{n_l+1}^l = x_{n_{l^*}+1}^{l^*} = \infty$.

We are interested in $m \geq 1$ future observations from each group. Here, we restrict attention to the case where the number of future observations is equal to the number of data observations. There are $\binom{2n_l}{n_l}$ orderings of the n_l future observations among the n_l data observations per group, and all possible orderings are equally likely. There are $\binom{2n_{l^*}}{n_{l^*}}$ orderings of the n_{l^*} future observations among the n_{l^*} data observations, and all possible orderings are equally likely. We consider all combinations of these possible orderings, denoted by O_ℓ for $\ell = 1, 2, ..., \binom{2n_{l^*}}{n_{l^*}} \prod_{l \neq l^*} \binom{2n_l}{n_l}$.

For each combination of orderings, O_ℓ, the corresponding Mosteller test statistic is denoted by r_ℓ. As the future observations are not precise, but only their number in each of the intervals of the partition created by the original data observations for their groups are known for a given ordering, we cannot calculate a precise value of r_ℓ related to a specific combination of orderings, but we can derive the minimum and maximum possible values; these are denoted by $\underline{r}_\ell$ and $\overline{r}_\ell$, respectively.

Let a specific ordering O_ℓ of n_l future observations among the n_l data observations be denoted by $(S_1^l, ..., S_{n_l+1}^l)$, with $S_{i_l}^l$ non-negative integers with $\sum_{i_l=1}^{n_l+1} S_{i_l}^l = n_l$. Let a specific ordering of n_{l^*} future observations among the n_{l^*} data observations be denoted by $(S_1^{l^*}, ..., S_{n_{l^*}+1}^{l^*})$, with $S_{i_{l^*}}^{l^*}$ non-negative integers with $\sum_{i_{l^*}=1}^{n_{l^*}+1} S_{i_{l^*}}^{l^*} = n_{l^*}$.

Now, let $a_l = \min\{i_l : S_{i_l}^l \neq 0, l \neq l^*\}$ be the index of the first interval from group l ($l \neq l^*$) that has at least one future observation, for $i_l = 1, 2, \ldots, n_l + 1$. Let, for $i_{l^*} = 1, 2, \ldots, n_{l^*} + 1$,

$$\kappa_1 = \max\{i_{l^*} : x_{i_{l^*}}^{l^*} < \min_{l \neq l^*} x_{a_l-1}^l\} \text{ and } \kappa_2 = \max\{i_{l^*} : x_{i_{l^*}-1}^{l^*} < \min_{l \neq l^*} x_{a_l}^l\}$$

To derive the minimum value r for a particular ordering combination O_ℓ, denoted by $\underline{r}_\ell$, all

$S^{l^*}_{i_{l^*}}$ future observations in the interval $(x^{l^*}_{i_{l^*}-1}, x^{l^*}_{i_{l^*}})$, $i_{l^*} = 1, 2, \ldots, n_{l^*} + 1$ are put at $x^{l^*}_{i_{l^*}}$, all $S^{l}_{i_l}$ future observations in the interval $(x^{l}_{i_l-1}, x^{l}_{i_l})$, $i_l = 1, 2, \ldots, n_l + 1$, are put at $x^{l}_{i_l-1}$, where $l \neq l^*$, then

$$\underline{r}_\ell = \sum_{t=1}^{\kappa_1} S_t^{l^*} \tag{5.4}$$

To derive the maximum value of r for a particular ordering O_ℓ, denoted by $\overline{r}_\ell$, all $S^{l^*}_{i_{l^*}}$ future observations in the interval $(x^{l^*}_{i_{l^*}-1}, x^{l^*}_{i_{l^*}})$, $i_{l^*} = 1, 2, \ldots, n_{l^*} + 1$ are put at $x^{l^*}_{i_{l^*}-1}$, all $S^{l}_{i_l}$ future observations in the interval $(x^{l}_{i_l-1}, x^{l}_{i_l})$, $i_l = 1, 2, \ldots, n_l + 1$, are put at $x^{l}_{i_l}$, where $l \neq l^*$, then

$$\overline{r}_\ell = \sum_{t=1}^{\kappa_2} S_t^{l^*} \tag{5.5}$$

Note, the ℓ is omitted from the right hand side for simplicity of notation. So, the probability $P_{(r)}$ in Equation (5.1), for equal sample sizes can be written in term of $\underline{r}_\ell$ and $\overline{r}_\ell$, as follows

$$P_{(\underline{r}_\ell)} = \frac{k(n!)(kn - \underline{r}_\ell)!}{(kn)!(n - \underline{r}_\ell)!} \tag{5.6}$$

$$P_{(\overline{r}_\ell)} = \frac{k(n!)(kn - \overline{r}_\ell)!}{(kn)!(n - \overline{r}_\ell)!} \tag{5.7}$$

The probability in Equation (5.3), when samples are of unequal size, can be written in term of $\underline{r}_\ell$ and $\overline{r}_\ell$ as follows,

$$P_{(\underline{r}_\ell)} = \frac{\sum_i n_i^{(\underline{r}_\ell)}}{N^{(\underline{r}_\ell)}} \tag{5.8}$$

$$P_{(\overline{r}_\ell)} = \frac{\sum_i n_i^{(\overline{r}_\ell)}}{N^{(\overline{r}_\ell)}} \tag{5.9}$$

The NPI lower reproducibility probability if the original test conclusion is rejection of H_0, is derived by counting the combinations, for which the same group as the original test has slipped to the left and $P(\underline{r}_\ell) \leq \alpha$. The corresponding NPI upper reproducibility probability is derived by counting the combinations for which the same group as the original test has slipped to the left and $P(\overline{r}_\ell) \leq \alpha$. Thus, the NPI lower and upper reproducibility probabilities are

$$\underline{RP} = \frac{1}{h} \sum_{\ell=1}^{h} 1\{P_{(\underline{r}_\ell)} \leq \alpha\} \tag{5.10}$$

$$\overline{RP} = \frac{1}{h} \sum_{\ell=1}^{h} 1\{P_{(\overline{r}_\ell)} \leq \alpha\} \tag{5.11}$$

where $h = \binom{2n_{l^*}}{n_{l^*}} \prod_{l \neq l^*} \binom{2n_l}{n_l}$ and $\ell = 1, 2, ..., h$. $1\{A\}$ is an indicator function which is equal to 1 if the event A occurs and 0 otherwise. This method to derive the NPI lower and upper reproducibility probabilities for the Mosteller test is suitable for small sample sizes and a limited number of groups due to computational limitations. In the case of large sample sizes and more

groups, we can apply the sampling of orderings and NPI bootstrap to derive approximations for the lower and upper reproducibility probabilities, which were introduced in Sections 1.5.2 and 1.5.3.

The exact NPI lower and upper reproducibility probabilities for the k groups Mosteller test can only be applied when the null hypothesis is rejected, that is one group has slipped to the left. Computational difficulties prevent deriving exact theoretical results for reproducibility probability for the Mosteller test when the null hypothesis is not rejected. However, deriving the exact lower and upper RP for the case of non-rejection of the null hypothesis is of interest as a topic of future research. In this chapter, the reproducibility for the Mosteller test for both cases rejection and non-rejection of the null hypothesis regardless of the group that has slipped in the future test will be consider through the NPI-RP-B method to approximate the NPI-RP. The methodology of this section will be applied in Section 5.8.

5.6 NPI-RP-SO for the Mosteller test

The NPI-RP-E approach, introduced in Section 5.5, for reproducibility of the Mosteller test can be implemented relatively for small sample sizes. For large sample sizes, the NPI reproducibility sampling of orderings (NPI-RP-SO) can be applied to estimat the NPI lower and upper reproducibility probabilities. The concept of the NPI-RP-SO has been introduced in Section 1.5.2, where we randomly sample r^* orderings from $\binom{2n}{n}$ possible orderings of the future observations among the data observations from each group [74, 75]. On the r^* orderings sampled, Equations (5.4) and (5.5) are applied to calculate the minimum and the maximum for the Mosteller test statistic. Suppose that H_0 is rejected in the original test, then the NPI lower reproducibility probability using the NPI-RP-SO is computed by counting the number of orderings for which $P(\underline{r}_\ell) \leq \alpha$ and the same group as the original test has slipped, divided by the number of orderings sampled, r^*. The NPI upper reproducibility probability is obtained by counting the number of orderings for which $P(\overline{r}_\ell) \leq \alpha$ and the same group as the original test has slipped, divided by r^*:

$$\widehat{\underline{RP}} = \frac{1}{r^*} \sum_{\ell=1}^{r^*} 1\{P_{(\underline{r}_\ell)} \leq \alpha\} \tag{5.12}$$

$$\widehat{\overline{RP}} = \frac{1}{r^*} \sum_{\ell=1}^{r^*} 1\{P_{(\overline{r}_\ell)} \leq \alpha\} \tag{5.13}$$

The NPI-RP-SO method will be illustrated via examples in Section 5.8, to investigate the NPI strong reproducibility probability for the Mosteller test.

5.7 NPI-RP-B for the Mosteller test

This section introduces the reproducibility probability for the Mosteller test, using the NPI-RP-B method introduced in Section 1.5.3. As mentioned earlier, when dealing with large sample sizes and an increasing number of groups, the number of orderings of the future observations among the data observations increases significantly. This leads to computational challenges that prevent the NPI-RP-E approach. BinHimd [18] proposed the use of the NPI-RP-B method as a heuristic method to approximate the reproducibility probability, as it avoids the complex calculations required by the NPI-RP-E approach. NPI-RP-B uses the point estimate for the NPI reproducibility probability instead of lower and upper reproducibility probabilities.

The application of Algorithm 2, which was introduced in Section 1.5.3 is adopted. The inputs into Algorithm 2 are the k original samples, their corresponding sample sizes, the number of runs T and the number of bootstrapped samples per run B. Summary statistics including the minimum, mean, median, maximum, of $RP_1, RP_2, ..., RP_T$ were calculated. Algorithm 2, will be implemented with both finite and infinite intervals, using the ranges introduced in Section 2.3. Approach I will be applied for finite interval, where the lower limit is taken to be the smallest value of the group minus the maximal distance between consecutive points, and the upper limit is taken to be equal to the largest value of the group plus the maximal distance between consecutive points. For infinite interval, Approach II will be used, which involves assuming the tail of a Normal distribution for data on the real line $(-\infty, \infty)$ and the tail of an Exponential distribution for data on $[0, \infty)$.

Section 5.9 presents the results of simulation studies using the NPI-RP-B approach for different scenarios, such as simulation under H_0 and H_1, with varying sample sizes and number of groups. In Section 5.9, the NPI-RP-B is used to study reproducibility and strong reproducibility for the Mosteller test. In the following section, the NPI-RP-B approach is also considered, to investigate whether or not the NPI-RP-B approach tends to provide values within the lower and upper NPI-RP-E and NPI-RP-SO.

5.8 Examples

This section studies the strong reproducibility probability for the Mosteller test. In Example 5.1, artificial data sets of ranks are used to study RP for the Mosteller test using the NPI-RP-E, NPI-RP-SO and NPI-RP-B approaches, as explained in Sections 5.5, 5.6 and 5.7. The NPI-RP-SO is considered for data sets from the literature in Examples 5.2, 5.3 and 5.4. The results of additional example to investigate the NPI-RP for the Mosteller test using the NPI-RP-SO are

n	Ranks		Test conclusion				NPI-RP-E		NPI-RP-B				NPI-RP-SO	
	X	Y	r	$P_{(r)}$	l^*	H_0	$\underline{RP}$	$\overline{RP}$	Min	Mean	Median	Max	$\widehat{\underline{RP}}$	$\widehat{\overline{RP}}$
5	1,2,3,4,5	6,7,8,9,10	5	0.008	X	R	0.389	1	0.926	0.944	0.945	0.960	0.391	1
	1,2,3,4,8	5,6,7,9,10	4	0.048	X	R	0.257	0.796	0.504	0.542	0.542	0.579	0.245	0.797
6	1,2,3,4,5,6	7,8,9,10,11,12	6	0.002	X	R	0.386	1	0.925	0.942	0.942	0.957	0.386	1
	1,2,3,4,5,8	6,7,9,10,11,12	5	0.015	X	R	0.275	0.824	0.536	0.575	0.574	0.613	0.276	0.821

Table 5.1: RP for the Mosteller test with $k = 2$, $\alpha = 0.05$

presented in the Appendix D.1.

Example 5.1. In this example a comparison of the three methods, NPI-RP-E, NPI-RP-SO and NPI-RP-B is carried out to study the strong reproducibility probability for the Mosteller test. In Tables 5.1 and 5.2, artificial data sets of ranks with equal samples sizes for $k = 2, 3$ groups are considered. The null hypothesis is that all groups are equal and the alternative hypothesis is that one group has slipped to the left. The original test conclusion is obtained and the smallest group has been identified. The null hypothesis is rejected if $P_{(r)} \leq \alpha$, where $\alpha = 0.05$.

The NPI-RP-E approach introduced in Section 5.5 is applied. For the ranks in Table 5.1, with $n_x = n_y = 5$, there are $\binom{10}{5} = 252$ possible orderings of 5 future observations among 5 data observations per group. So, there are $\binom{10}{5}^2 = 63504$ orderings combinations to consider in the calculation of $\underline{RP}$ and $\overline{RP}$. With $n_x = n_y = 6$, there are $\binom{12}{6}^2 = 853776$ orderings combinations each $\underline{RP}$ and $\overline{RP}$ value is based on.

The NPI-RP results in Tables 5.1 and 5.2 are introduced for the cases when the null hypothesis is rejected, as the NPI-RP-E approach introduced in Section 5.5 requires known the group that has slipped to the left. The results in Tables 5.1 and 5.2 show that the $\underline{\text{RP}}$ is low and below 0.5 for all cases. In Table 5.2, with $n = 3$, for the ranks in the first line, the lower is 0.125, the reason for that is the maximum future X ranks is less than 3 with probability 0.5, the minimum future Y ranks is greater than 4 with probability 0.5 and the minimum future Z ranks is greater than 7 with probability 0.5. Similarly, when $n = 4$ and the data are perfectly ordered for the ranks in the third line, the lower RP is 0.125, this is because the maximum future X ranks is less than 4 with probability 0.5, the minimum future Y ranks is greater than 5 with probability 0.5 and the minimum future Z ranks is greater than 9 with probability 0.5. As these individual events happen with probability 0.5 in the NPI framework, with the independence between the three groups leads to the lower reproducibility probability $0.5 \times 0.5 \times 0.5 = 0.125$.

Here, we also investigate whether or not the NPI-RP-B method tends to provide values within the lower and upper NPI-RP-E and NPI-RP- SO. The NPI-RP-SO approach is considered with the number of orderings sampled $r^* = 2000$. The NPI-RP-B method uses Algorithm 2.

n	Ranks			Test conclusion				NPI-RP-E		NPI-RP-B				NPI-RP-SO	
	X	Y	Z	r	$P_{(r)}$	l^*	H_0	$\underline{RP}$	$\overline{RP}$	Min	Mean	Median	Max	$\widehat{\underline{RP}}$	$\widehat{\overline{RP}}$
3	1,2,3	4,5,6	7,8,9	3	0.036	X	R	0.125	1	0.815	0.848	0.850	0.872	0.122	1
	6,7,9	4,5,8	1,2,3	3	0.036	Z	R	0.125	1	0.640	0.672	0.673	0.698	0.126	1
4	1,2,3,4	5,6,7,8	9,10,11,12	4	0.006	X	R	0.125	1	0.813	0.845	0.845	0.870	0.122	1
	1,2,3,4	5,6,7,10	8,9,11,12	4	0.006	X	R	0.125	1	0.642	0.674	0.675	0.716	0.125	1

Table 5.2: RP for the Mosteller test with $k = 3$, $\alpha = 0.05$

Algorithm 2 is performed on finite intervals using Approach I, with $B = 1000$ and $T = 100$. The minimum, mean, median and maximum of $RP_1, \dots RP_{100}$, were calculated. The mean of $RP_1, \dots RP_{100}$ is the NPI-RP-B value. It can be inferred that the NPI-RP-B estimates are all within the NPI lower and upper reproducibility probabilities derived by the NPI-RP-E and the NPI-RP-SO methods. This agrees with the results of the umbrella alternatives tests in Chapter 4, and other NPI studies of test reproducibility [2, 18, 97]. The minimum and the maximum values of the RP using the NPI-RP-B are very close, while the lower and upper RP using the NPI-RP-E and NPI-RP-SO methods are very different. For example, for ranks in the first line in Table 5.1, the different between the minimum and the maximum is equal to 0.034, while between $\underline{RP}$ and $\overline{RP}$ is 0.611. The reason is that for the lower and upper RP, we move the probability masses to the extremes, while, in the NPI-RP-B method we sample future observations between the intervals. Assigning the probability masses to the extremes also leads to very small lower RP because we make it very pessimistic for the event to occur again.

Example 5.2. In this example, we are examining the NPI strong reproducibility for the Mosteller test using the NPI-RP-E and NPI-RP-SO approaches. We are using the data set from Table 5.3, which shows the extent of coffee berry disease. This data set includes the percentage of infections in test berries for farms that were not sprayed (X) and those that were sprayed (at least 14 months prior to sampling) with a fungicide (Y) [108]. Each group consists of 7 observations. We address tied observations by adding a small amount, as detailed in Section 1.4. The data is visualized in Figure 5.1.

The null hypothesis is that both groups are identical and the alternative hypothesis is that one group has slipped to the left. The level of significant is $\alpha = 0.05$. Applying the Mosteller test, group X is the group that has slipped to the left, with $r = 5$. The probability of obtaining 5 observations from group X that are less than all observations from group Y, using Equation (5.1), is $P_{(r)} = 0.023$ which is less than 0.05. So, the null hypothesis is rejected.

Applying the NPI-RP-E approach introduced in Section 5.5, there are $\binom{14}{7}\binom{14}{7} = 11778624$ orderings combinations to consider in the calculation of the NPI lower and upper reproducibility

Unsprayed (X)	0.75	1.76	2.48	4.88	5.10	6.01	7.13
Sprayed (Y)	5.68	5.69	11.63	16.30	21.46	33.30	44.20

Table 5.3: Percentage infections in test berries, for Example 5.2

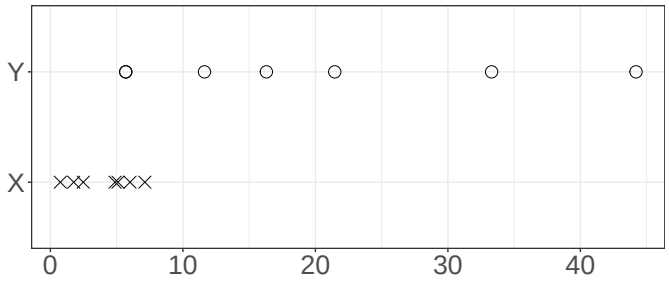


Figure 5.1: Visualization of the percentage infections in test berries data, for Example 5.2

probabilities. We obtain the exact values $\underline{RP} = 0.289$ and $\overline{RP} = 0.785$. The NPI lower reproducibility probability is low because the $P_{(r)}$ value is close to the level of significant $\alpha = 0.05$. To apply the NPI-RP-SO method, we sampled different number of orderings r^* to calculate approximations for the NPI lower and upper reproducibility probabilities. The 95% confidence interval was computed for both lower and upper reproducibility probabilities, as introduced in Section 1.5.2. Based on the NPI-RP-SO results in Table 5.4, an accurate approximations of the NPI lower and upper reproducibility probabilities can be obtained by considering the number orderings sampled equal or greater than 2,000.

Example 5.3. This example considers the NPI strong reproducibility for the Mosteller test with $k = 2$ and unequal sample sizes, using the NPI-RP-SO approach. The data set given in Table 5.5, for the average delay times for subjects with Parkinsonism disease (X) and normal subjects (Y), in performing fast tasks is used [108]. Figure 5.2 displays visualization of the data.

The null hypothesis is that both groups are identical and the alternative hypothesis is that one group has slipped to the left. The level of significant is $\alpha = 0.05$. Applying the Mosteller test with $n_x = 10$ and $n_y = 8$, group X is the group that has slipped to the left, with $r = 8$. The probability of obtaining 8 observations from group X that are less than all observations from group Y, using Equation (5.1), is $P_{(r)} = 0.001$ which is less than 0.05. So, the null hypothesis is rejected.

In the NPI-RP-E approach introduced in Section 5.5, for $n_x = 10$, there are $\binom{20}{10} = 184756$ possible orderings and for $n_y = 8$, there are $\binom{16}{8} = 12870$ possible orderings. So, there are $\binom{20}{10}\binom{16}{8} = 2377809720$ orderings combinations to be considered in the calculation of the NPI

r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.300	(0.210, 0.390)	0.750	(0.665, 0.835)
500	0.272	(0.233, 0.311)	0.740	(0.702, 0.778)
1,000	0.271	(0.243, 0.299)	0.770	(0.744, 0.796)
2,000	0.287	(0.267, 0.307)	0.771	(0.753, 0.789)
5,000	0.281	(0.269, 0.294)	0.787	(0.775, 0.798)
10,000	0.289	(0.280, 0.298)	0.784	(0.776, 0.792)
50,000	0.290	(0.286, 0.294)	0.785	(0.781, 0.789)
100,000	0.291	(0.288, 0.294)	0.787	(0.784, 0.790)
150,000	0.290	(0.288, 0.292)	0.785	(0.783, 0.787)

Table 5.4: NPI-RP-SO for the Mosteller test with $k = 2$ and $n = 7$, for Example 5.2

Normal (X)	206	211	213	229	258	267	281	281	317	321
Disease (Y)	290	290	360	400	403	420	460	660		

Table 5.5: The average time delay of tasks, for Example 5.3

lower and upper reproducibility probabilities using the NPI-RP-E approach. This number of combinations is slightly large, however, the NPI-RP-E results can be obtained using high performance computer. The NPI-RP-SO approach is applied, as introduced in Section 5.6. We sampled different number of orderings r^* to calculate approximations for $\underline{RP}$ and $\overline{RP}$. The 95% confidence interval was computed for both $\widehat{\underline{RP}}$ and $\widehat{\overline{RP}}$. From the NPI-RP results presented in Table 5.6, it can be concluded that reasonable approximations of the NPI lower and upper reproducibility probabilities for the Mosteller test, can be obtained by considering the number orderings sampled equal or greater than 2,000 which is a quite small number when compared with the number of all possible orderings.

Example 5.4. This example investigates NPI strong reproducibility for the Mosteller test via the NPI-RP-SO approach for $k = 4$ groups, using the data given in Table 5.7. This data consists of verbal IQ scores for four groups of first grade children, each group containing 23 children, residing in four different types of communities [99]. Figure 5.3 displays visualization of the data. The `jitter` function in R is used to break ties in the data. Applying the Mosteller test, the very isolated group is the group that has slipped to the left, with $r = 13$. The probability of obtaining 13 observations in this group that are less than all observations in the other groups, using Equation (5.1), is $P_{(r)} = 2.050 \times 10^{-9}$, which is less than $\alpha = 0.05$, indicating evidence

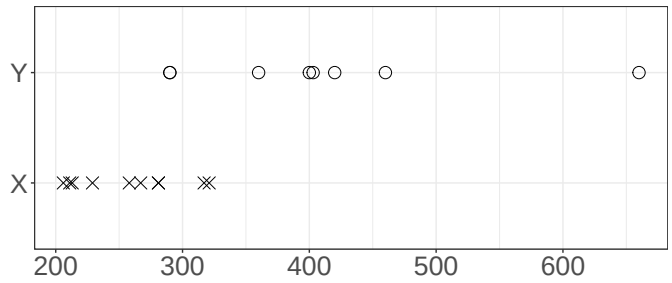


Figure 5.2: Visualization of the average time delay of tasks data, for Example 5.3

r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.450	(0.352, 0.548)	0.980	(0.930, 0.998)
500	0.464	(0.420, 0.508)	0.978	(0.965, 0.991)
1,000	0.462	(0.431, 0.493)	0.980	(0.971, 0.989)
2,000	0.468	(0.446, 0.490)	0.976	(0.969, 0.983)
5,000	0.461	(0.448, 0.475)	0.976	(0.972, 0.981)
10,000	0.460	(0.450, 0.469)	0.980	(0.977, 0.983)
50,000	0.464	(0.460, 0.468)	0.977	(0.460, 0.468)
100,000	0.468	(0.465, 0.471)	0.978	(0.977, 0.979)
150,000	0.466	(0.463, 0.469)	0.979	(0.978, 0.980)

Table 5.6: NPI-RP-SO for the Mosteller test with $k = 2$, $n_x = 10$ and $n_y = 8$, for Example 5.3

against H_0.

In the NPI approach, there are $\binom{46}{23}\binom{46}{23}\binom{46}{23}\binom{46}{23} = 4.595 \times 10^{51}$ orderings combinations to consider in the calculation of the NPI lower and upper reproducibility probabilities using the NPI-RP-E approach, and it is unfeasible to go through this large number of orderings. Thus, in Table 5.8, the NPI-RP-SO method is applied with different number of orderings sampled r^*, to calculate approximation values of NPI lower and upper reproducibility probabilities. The 95% confidence intervals for $\widehat{\underline{RP}}$ and $\widehat{\overline{RP}}$ are also obtained as in Table 5.8. It can be concluded that good approximations of the NPI lower and upper reproducibility probabilities for the Mosteller test can be achieved when the number of sampled orderings, r^*, is 10,000 or more.

Very isolated (X)	20	20	21	21	21	22	22	22	22	23	23	23
	23	25	25	26	27	28	28	30	30	37	44	
Moderately isolated (Y)	26	26	26	27	28	29	29	29	32	33	33	33
	33	34	34	34	34	35	36	40	42	44	45	
Rural nonisolated (Z)	26	27	30	30	30	33	33	33	35	35	35	35
	36	36	36	37	37	39	39	39	40	41	42	
Urban ghetto (V)	24	25	25	28	32	34	36	36	37	37	38	40
	40	41	41	42	42	42	43	43	44	45	45	

Table 5.7: Verbal IQ scores data, for Example 5.4

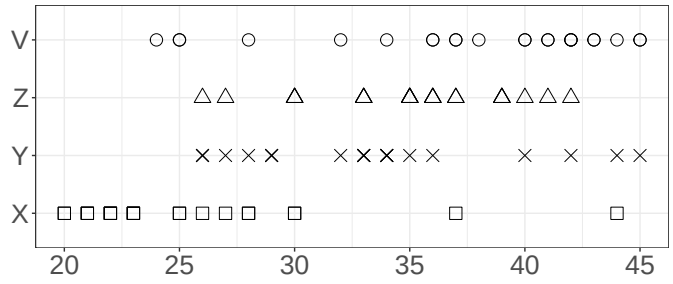


Figure 5.3: Visualization of the IQ scores data, for Example 5.4

r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.120	(0.056, 0.184)	1	(0.964, 1)
500	0.150	(0.119, 0.181)	1	(0.993, 1)
1,000	0.127	(0.106, 0.148)	1	(0.996, 1)
5,000	0.124	(0.115, 0.133)	0.999	(0.998, 1.000)
10,000	0.123	(0.117, 0.129)	0.999	(0.999, 1.000)
50,000	0.126	(0.123, 0.129)	0.999	(0.999, 1.000)
100,000	0.126	(0.124, 0.128)	1.000	(0.999, 1.000)
150,000	0.126	(0.124, 0.128)	0.999	(0.999, 1.000)

Table 5.8: NPI-RP-SO for the Mosteller test with $k = 4$ and $n = 23$, for Example 5.4

Case	k	Simulation
1	3	$N(0,1)$
2	3	$X \sim N(0,1)$, $Y \sim N(1.5,1)$, $Z \sim N(2,1)$
3	5	$N(0,1)$
4	5	$X \sim N(0,1)$, $Y \sim N(1,1)$, $Z \sim N(1.5,1)$, $V \sim N(2,1)$, $W \sim N(2.5,1)$

Table 5.9: Simulation cases for the Mosteller test

5.9 Simulation study

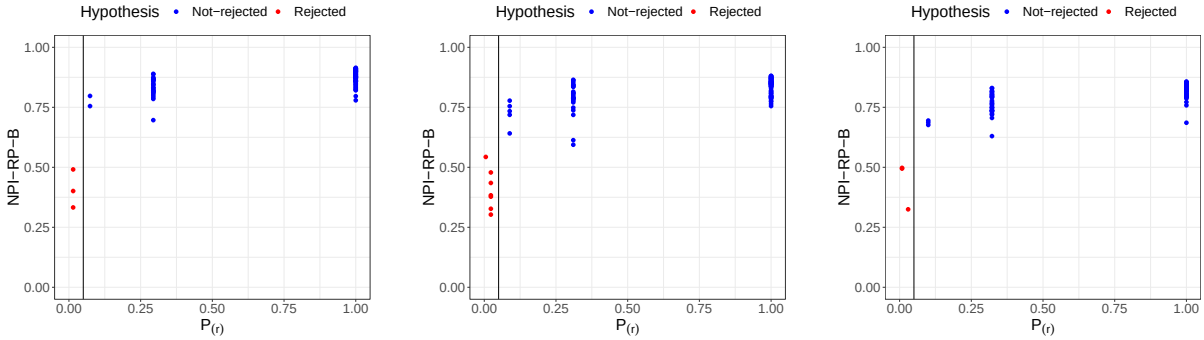
This section studies reproducibility and strong reproducibility probability for the Mosteller test, via simulations using Algorithm 2. The NPI-RP-B method is preformed with infinite support using Approach II, introduced in Section 5.7, which involves assuming the tail of a Normal distribution when considering real-valued data sets. We test the null hypothesis that all groups are equal against the alternative hypothesis that one group has slipped to the left. The level of significance is $\alpha = 0.05$. The decision rule for this test is to reject H_0, if $P_{(r)}$ obtained from Equation (5.1) is less than $\alpha = 0.05$. Data were simulated under H_0 and H_1, as presented in Table 5.9. To study the impact of the number of groups and the sample size on the reproducibility probability, the simulation is considered with the number of groups $k = 3$, 5 and each case introduced in Table 5.9 is considered with the sample sizes $n = 6$, 10, 20.

The inputs for the simulation study in Tables 5.10 through Table 5.21 are as follows: Algorithm 2 is applied with $B = 1000$ and $T = 100$. For each run, one sample of size n is generated from each of the distributions given in the Table 5.9, the Mosteller test is performed on the these samples, and the tests outcomes are obtained and the reproducibility estimates for the Mosteller test are calculated. In each table, the reproducibility probability estimates have been reported for 10 simulated data sets. For the same value $P_{(r)}$, the reproducibility probability estimates differs from one data set to another data set. These small variations in the RP estimates are due to variations in the original samples and in the NPI-B samples.

The relationship between NPI-RP-B and $P_{(r)}$ for the Mosteller test is examined in the simulations. The observed $P_{(r)}$ and the NPI-RP-B estimates for 100 data sets are displayed in Figures 5.4, 5.5, 5.6 and 5.7. Note that, the level of significance $\alpha = 0.05$ is represented on the figures by a vertical line. Based on these figures, it is clear that the NPI-RP-B estimates are low when $P_{(r)}$ is close to the threshold 0.05, which is as expected. The NPI-RP-B estimates also tend to be lower in the case of rejection than for non-rejection. A similar pattern has been observed in the previous chapters in the investigation of the NPI reproducibility for the

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.893	0.913	0.914	0.934
2	1	1	NR	0.812	0.841	0.842	0.871
3	1	1	NR	0.851	0.872	0.871	0.905
4	1	1	NR	0.870	0.895	0.896	0.920
5	2	0.294	NR	0.810	0.831	0.829	0.859
6	2	0.294	NR	0.794	0.825	0.826	0.847
7	2	0.294	NR	0.805	0.833	0.833	0.860
8	3	0.074	NR	0.679	0.717	0.719	0.746
9	3	0.074	NR	0.701	0.737	0.735	0.773
10	4	0.015	R	0.463	0.499	0.498	0.537

Table 5.10: RP for the Mosteller test, with Case 1, $n = 6$, $\alpha = 0.05$



(a) $n = 6$ (b) $n = 10$ (c) $n = 20$

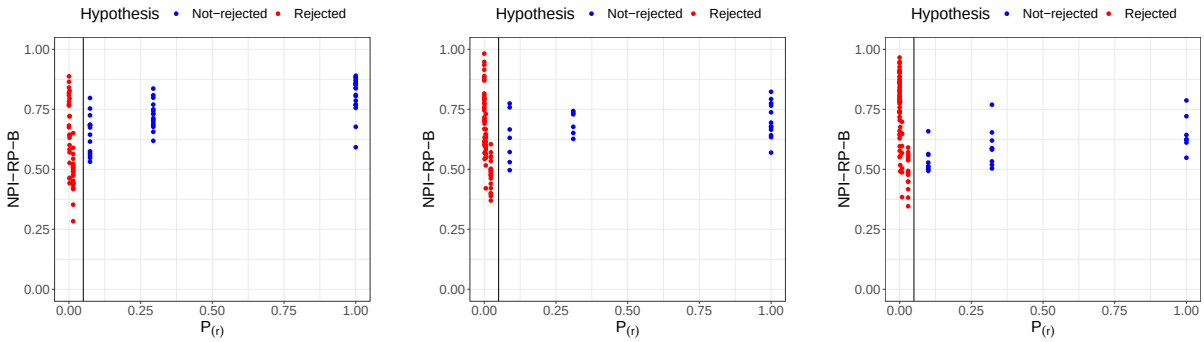
Figure 5.4: NPI-RP-B for the Mosteller test under H_0, with Case 1, $\alpha = 0.05$

Jonckheere-Terpstra test in Chapter 3 and the umbrella alternatives tests in Chapter 4, and with other NPI-RP applications for other tests [2, 18, 97]. Further simulations were performed for data generated under H_0 and H_1 for $k = 3$ and unequal sample sizes, the results are presented in the Appendix D.2.

In Tables 5.22, 5.23 and 5.24, the NPI strong reproducibility probability for the Mosteller test is also explored. Data were generated under H_0 and H_1 with $k = 3, 4$ groups and the sample size $n = 10$. For $k = 3$ and under H_0, Case 1 in Table 5.9 is considered. Under H_1, original data were generated from Normal distribution with different means $\mu_x = 0$, $\mu_y = 0.5$ and $\mu_z = 1.5$, and standard deviation 1. For $k = 4$, original data were generated from Normal distribution with different means $\mu_x = 0$, $\mu_y = 0.5$, $\mu_z = 1.5$ and $\mu_v = 2$, with standard deviation 1. Algorithm 2 is implemented with $B = 10000$ and $T = 1$. From the results in Tables 5.22,

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.831	0.867	0.866	0.897
2	1	1	NR	0.833	0.875	0.877	0.906
3	1	1	NR	0.790	0.823	0.824	0.859
4	1	1	NR	0.840	0.869	0.869	0.892
5	2	0.310	NR	0.804	0.838	0.839	0.866
6	2	0.310	NR	0.741	0.768	0.767	0.803
7	2	0.310	NR	0.767	0.798	0.798	0.824
8	3	0.089	NR	0.679	0.708	0.708	0.742
9	3	0.089	NR	0.651	0.690	0.689	0.736
10	4	0.023	R	0.265	0.311	0.310	0.350

Table 5.11: RP for the Mosteller test, with Case 1, $n = 10$, $\alpha = 0.05$



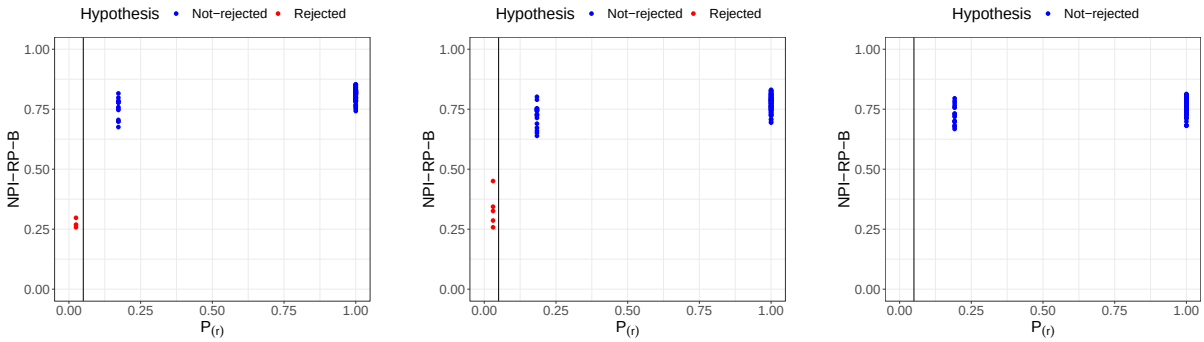
(a) $n = 6$ (b) $n = 10$ (c) $n = 20$

Figure 5.5: NPI-RP-B for the Mosteller test under H_1, with Case 2, $\alpha = 0.05$

5.23 and 5.24, it can be concluded that for the rejection cases with one group has slipped to the left, the rejection is reproduced in the future bootstrapped samples with the same group contributes the most in the NPI-RP. As shown for the original samples 6-10 in Table 5.23 and 3-10 in Table 5.24, the original test conclusion is the rejection of the null hypothesis with group X has slipped. Here, the rejection is reproduced largely with the same group X has slipped in the future samples.

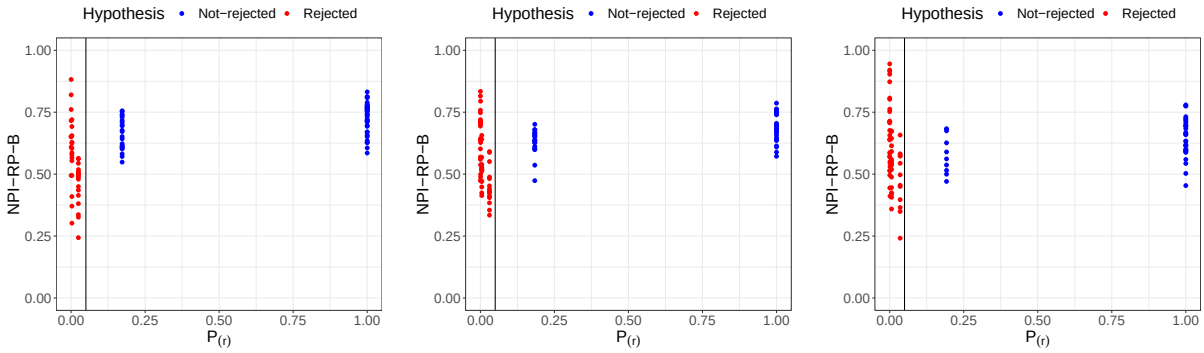
Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.829	0.854	0.853	0.880
2	1	1	NR	0.821	0.847	0.847	0.879
3	1	1	NR	0.804	0.837	0.839	0.869
4	1	1	NR	0.779	0.812	0.812	0.835
5	2	0.322	NR	0.779	0.817	0.818	0.846
6	2	0.322	NR	0.795	0.831	0.832	0.857
7	2	0.322	NR	0.746	0.784	0.784	0.817
8	2	0.322	NR	0.709	0.747	0.747	0.781
9	3	0.100	NR	0.712	0.749	0.750	0.781
10	3	0.100	NR	0.572	0.614	0.612	0.648

Table 5.12: RP for the Mosteller test, with Case 1, $n = 20$, $\alpha = 0.05$



(a) $n = 6$ (b) $n = 10$ (c) $n = 20$

Figure 5.6: NPI-RP-B for the Mosteller test under H_0, with Case 3, $\alpha = 0.05$



(a) $n = 6$ (b) $n = 10$ (c) $n = 20$

Figure 5.7: NPI-RP-B for the Mosteller test under H_1, with Case 4, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.787	0.812	0.812	0.844
2	1	1	NR	0.761	0.786	0.787	0.815
3	2	0.294	NR	0.772	0.802	0.801	0.832
4	2	0.294	NR	0.758	0.796	0.796	0.828
5	2	0.294	NR	0.688	0.748	0.748	0.775
6	3	0.074	NR	0.724	0.760	0.759	0.794
7	3	0.074	NR	0.732	0.768	0.769	0.795
8	4	0.015	R	0.307	0.341	0.340	0.374
9	5	0.002	R	0.458	0.494	0.496	0.534
10	6	1.616×10^{-4}	R	0.693	0.727	0.727	0.756

Table 5.13: RP for the Mosteller test, with Case 2, $n = 6$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.766	0.795	0.794	0.829
2	1	1	NR	0.700	0.741	0.740	0.780
3	2	0.310	NR	0.553	0.600	0.599	0.659
4	3	0.089	NR	0.674	0.710	0.708	0.749
5	3	0.089	NR	0.536	0.564	0.564	0.613
6	4	0.023	R	0.364	0.417	0.419	0.447
7	4	0.023	R	0.480	0.517	0.518	0.563
8	5	0.005	R	0.515	0.554	0.554	0.600
9	6	0.001	R	0.679	0.714	0.716	0.746
10	8	2.307×10^{-5}	R	0.804	0.827	0.827	0.853

Table 5.14: RP for the Mosteller test, with Case 2, $n = 10$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.584	0.638	0.637	0.673
2	3	0.100	NR	0.569	0.605	0.606	0.636
3	4	0.030	R	0.390	0.434	0.433	0.484
4	5	0.009	R	0.530	0.564	0.565	0.604
5	6	0.002	R	0.624	0.662	0.662	0.700
6	7	6.022×10^{-4}	R	0.644	0.689	0.690	0.731
7	8	1.477×10^{-4}	R	0.742	0.769	0.768	0.806
8	9	3.408×10^{-5}	R	0.808	0.842	0.841	0.868
9	11	1.470×10^{-6}	R	0.894	0.927	0.928	0.954
10	15	8.744×10^{-10}	R	0.947	0.963	0.963	0.977

Table 5.15: RP for the Mosteller test, with Case 2, $n = 20$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.807	0.839	0.839	0.868
2	1	1	NR	0.805	0.835	0.836	0.860
3	1	1	NR	0.778	0.825	0.826	0.868
4	1	1	NR	0.771	0.798	0.797	0.835
5	2	0.172	NR	0.756	0.789	0.788	0.828
6	2	0.172	NR	0.725	0.765	0.766	0.793
7	2	0.172	NR	0.707	0.745	0.744	0.772
8	2	0.172	NR	0.736	0.763	0.763	0.794
9	3	0.025	R	0.321	0.352	0.352	0.398
10	3	0.025	R	0.338	0.376	0.376	0.412

Table 5.16: RP for the Mosteller test, with Case 3, $n = 6$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.804	0.832	0.832	0.857
2	1	1	NR	0.784	0.809	0.810	0.835
3	1	1	NR	0.748	0.774	0.773	0.810
4	1	1	NR	0.716	0.760	0.760	0.798
5	1	1	NR	0.769	0.800	0.802	0.829
6	2	0.184	NR	0.665	0.711	0.712	0.740
7	2	0.184	NR	0.699	0.744	0.746	0.777
8	2	0.184	NR	0.668	0.714	0.715	0.743
9	2	0.184	NR	0.691	0.724	0.722	0.755
10	3	0.031	R	0.221	0.259	0.259	0.290

Table 5.17: RP for the Mosteller test, with Case 3, $n = 10$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.751	0.784	0.784	0.812
2	1	1	NR	0.769	0.796	0.795	0.825
3	1	1	NR	0.744	0.775	0.775	0.802
4	1	1	NR	0.750	0.780	0.780	0.812
5	1	1	NR	0.761	0.795	0.796	0.822
6	2	0.192	NR	0.731	0.770	0.770	0.799
7	2	0.192	NR	0.699	0.756	0.756	0.788
8	2	0.192	NR	0.747	0.778	0.778	0.813
9	2	0.192	NR	0.668	0.707	0.708	0.737
10	4	0.006	R	0.307	0.356	0.356	0.389

Table 5.18: RP for the Mosteller test, with Case 3, $n = 20$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.709	0.744	0.745	0.778
2	1	1	NR	0.660	0.693	0.691	0.732
3	2	0.172	NR	0.629	0.668	0.670	0.697
4	2	0.172	NR	0.565	0.604	0.603	0.644
5	3	0.025	R	0.293	0.321	0.319	0.358
6	3	0.025	R	0.424	0.462	0.462	0.503
7	4	0.003	R	0.387	0.424	0.424	0.456
8	4	0.003	R	0.472	0.513	0.512	0.552
9	4	0.003	R	0.507	0.539	0.540	0.580
10	5	2.105×10^{-4}	R	0.507	0.542	0.541	0.575

Table 5.19: RP for the Mosteller test, with Case 4, $n = 6$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.678	0.709	0.711	0.753
2	1	1	NR	0.657	0.705	0.705	0.753
3	2	0.184	NR	0.636	0.678	0.680	0.710
4	3	0.031	R	0.349	0.384	0.383	0.432
5	3	0.031	R	0.396	0.432	0.431	0.469
6	4	0.005	R	0.503	0.551	0.552	0.592
7	5	5.947×10^{-4}	R	0.570	0.608	0.609	0.645
8	6	6.608×10^{-5}	R	0.564	0.604	0.604	0.653
9	6	6.608×10^{-5}	R	0.667	0.704	0.704	0.738
10	7	6.007×10^{-6}	R	0.707	0.738	0.739	0.786

Table 5.20: RP for the Mosteller test, with Case 4, $n = 10$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.709	0.750	0.749	0.788
2	2	0.192	NR	0.566	0.606	0.605	0.655
3	3	0.035	R	0.306	0.355	0.356	0.394
4	4	0.006	R	0.507	0.548	0.550	0.582
5	4	0.006	R	0.529	0.580	0.582	0.608
6	5	0.001	R	0.593	0.633	0.636	0.661
7	6	1.626×10^{-4}	R	0.646	0.682	0.682	0.716
8	7	2.421×10^{-5}	R	0.766	0.798	0.799	0.825
9	7	2.421×10^{-5}	R	0.847	0.874	0.874	0.897
10	10	5.337×10^{-8}	R	0.851	0.878	0.878	0.912

Table 5.21: RP for the Mosteller test, with Case 4, $n = 20$, $\alpha = 0.05$

						Future l^*		
Samples	r	$P_{(r)}$	l^*	H_0	NPI-RP-B	X	Y	Z
1	1	1	X	NR	0.866	0.351	0.270	0.245
2	1	1	Z	NR	0.748	0.086	0.141	0.522
3	1	1	Z	NR	0.794	0.341	0.022	0.431
4	2	0.318	X	NR	0.770	0.561	0.172	0.037
5	2	0.310	X	NR	0.775	0.558	0.154	0.063
6	2	0.318	Z	NR	0.768	0.177	0.057	0.534
7	3	0.089	Z	NR	0.748	0.086	0.141	0.522
8	3	0.096	X	NR	0.711	0.487	0.049	0.175
9	4	0.023	Z	R	0.390	0.003	0.008	0.380
10	4	0.023	Z	R	0.496	0.003	0.002	0.491

Table 5.22: RP for the Mosteller test, with $k = 3$ and the original samples from $N(0, 1)$, $n = 10$, $\alpha = 0.05$

						Future l^*		
Samples	r	$P_{(r)}$	l^*	H_0	NPI-RP-B	X	Y	Z
1	1	1	X	NR	0.878	0.345	0.163	0.369
2	1	1	X	NR	0.846	0.382	0.290	0.175
3	2	0.310	X	NR	0.723	0.514	0.204	0.005
4	2	0.310	X	NR	0.625	0.375	0.209	0.041
5	3	0.089	X	NR	0.643	0.519	0.062	0.062
6	4	0.023	X	R	0.433	0.425	0.002	0.007
7	4	0.023	X	R	0.408	0.397	0.003	0.008
8	4	0.023	X	R	0.506	0.494	0.011	0.000
9	5	0.005	X	R	0.541	0.536	0.003	0.002
10	5	0.005	X	R	0.593	0.590	0.003	0.000

Table 5.23: RP for the Mosteller test, with $k = 3$ and $X \sim N(0, 1)$, $Y \sim N(0.5, 1)$ and $Z \sim N(1.5, 1)$, $n = 10$, $\alpha = 0.05$, $B = 10000$, $T = 1$.

						Future l^*			
Samples	r	$P_{(r)}$	l^*	H_0	NPI-RP-B	X	Y	Z	V
1	1	1	X	NR	0.676	0.231	0.378	0.059	0.007
2	2	0.231	X	NR	0.595	0.438	0.124	0.003	0.031
3	3	0.049	X	R	0.469	0.400	0.067	0.002	0.000
4	3	0.049	X	R	0.426	0.289	0.040	0.093	0.003
5	3	0.049	X	R	0.421	0.373	0.028	0.006	0.014
6	4	0.009	X	R	0.495	0.458	0.022	0.014	0.000
7	4	0.009	X	R	0.468	0.432	0.021	0.008	0.007
8	4	0.009	X	R	0.494	0.455	0.019	0.020	0.000
9	5	0.002	X	R	0.503	0.471	0.015	0.003	0.015
10	5	0.002	X	R	0.679	0.669	0.006	0.005	0.000

Table 5.24: RP for the Mosteller test, with $k = 4$ and $X \sim N(0, 1)$, $Y \sim N(0.5, 1)$, $Z \sim N(1.5, 1)$ and $V \sim N(2, 1)$, $n = 10$, $\alpha = 0.05$, $B = 10000$, $T = 1$.

5.10 Concluding remarks

This chapter explored the NPI reproducibility probability for the Mosteller test, for the alternative that one group has slipped to the left of the other groups. The exact NPI lower and upper reproducibility probabilities for the Mosteller test are derived for the scenario of rejection of the null hypothesis. As the sample size and the number of groups increase, the number of possible orderings of the future observations among the data observations increases significantly, leading to computational challenges using the exact NPI reproducibility probability (NPI-RP-E) approach. To this end, two NPI-based approaches are implemented, namely, the sampling of orderings (NPI-RP-SO) and the NPI-bootstrap (NPI-RP-B). The employment of the NPI-RP-SO and NPI-RP-B approaches have the advantage of avoiding the complexities involved in computations of the exact lower and upper bounds.

The NPI-RP-SO and NPI-RP-B methods has been applied to a variety of scenarios via simulation studies. The findings of the NPI reproducibility probability for the Mosteller test are consistent with previous NPI studies of test reproducibility [2, 18, 33, 97]. Notably, these studies show that the reproducibility is low close the the boundaries of the rejection region. The findings also show consistency with the results in Chapters 3 and 4 which show that the reproducibility is low when the p-values are close to the test threshold, and it is lower when the null hypothesis is rejected more than when it is not rejected. The reason for that is the presence of some sort of direction in the alternative.

For future research it is of interest to consider the reproducibility probability for other slippage tests, such as the slippage test proposed by Granger and Neave [52], for the alternative that one group has slipped from the other groups. The reproducibility for the slippage tests involving more than one group being slipped is a topic for future research. It is of interest to derive NPI-RP-E approach for Mosteller test that considers both cases when the null hypothesis is rejected and not rejected which may require developing some methods. Studying reproducibility probability for scale problem is also of interest for future research.

Chapter 6

Conclusions

This chapter summarises the main results of this thesis and concludes with some future research topics. In this thesis, the reproducibility probability was explored for statistical hypothesis tests using the NPI approach. It has been noted that there is no standardised definition of the reproducibility probability within the classical frequentist statistics framework. In the NPI setting, the reproducibility probability is considered from a prediction perspective. The main question that this thesis addresses is: if a statistical test were repeated, under the same circumstances, would it lead to the same conclusion with regard to rejection or non-rejection of the null hypothesis?

In Chapter 2, the NPI reproducibility probability was introduced for general alternatives tests, including the Kruskal Wallis test and the Analysis of Variance (ANOVA) test. The NPI bootstrap method was performed for different scenarios. The findings suggest that the reproducibility is low (close to 0.5) when the observed test statistic value is close to the test threshold. Reproducibility tends to increase when the test statistic moves away from the threshold. In principle for the general alternatives tests, if there are multiple groups, the reproducibility close to the rejection region is lower in the case of non-rejection than for rejection of the null hypothesis substantially less than 0.5. All different scenarios provide similar results for both the nonparametric Kruskal-Wallis test and the parametric ANOVA test. As a result it can be concluded that when the test has been performed and it satisfies the criteria of the power and the level of significance, the RP could also be quit low. It would be interesting for future research to derive closed formula for the lower and upper reproducibility probability for these tests using the NPI-RP-E approach.

In Chapter 3, the NPI reproducibility probability was explored for the Jonckheere-Terpstra (JT) test. The JT test is used for the alternative hypothesis that the location parameters are ordered in a specific way, such as an increasing or decreasing trend. The NPI-RP-B method was

adopted to investigate the reproducibility probability for the JT test for different scenarios. The findings in Chapter 3 indicate that when the test statistic value is close to the test threshold the NPI reproducibility do not provide strong evidence in favour of the reproducibility of the test results, particularly, in case of rejection than for non-rejection of the null hypothesis. The reason for that is the presence of direction or trend in the alternatives. Deriving the exact lower and upper reproducibility probabilities for the JT test is of interest for future work.

Chapter 4 investigates the NPI reproducibility for the umbrella alternatives tests, namely the Mack-Wolf (MW) test and Esra-Fikri (EF) test. The exact lower and upper reproducibility probabilities for the MW test are derived for three groups; however, for more than three groups and large samples the application of the exact method becomes computationally challenging. In this case, two NPI-based approaches are implemented, namely, the sampling of orderings and the NPI-bootstrap technique. The results of this chapter indicate that the reproducibility is low when the test statistic value is close to the rejection threshold, and tends to be lower in the case of rejection than for non-rejection. Finding the exact lower and upper reproducibility probabilities for the MW test for more than three groups is of interest for future research. The reproducibility for other umbrella alternatives tests can also be studied.

In Chapter 5, the NPI reproducibility probability is investigated for the Mosteller test. Mosteller test is used to determine whether one group has slipped to the right or slipped to the left. The NPI bootstrap approach was implemented and it was observed that the results in this chapter are consistent with the results in Chapters 3 and 4, as these tests are used for directional alternatives. It would be interesting to study NPI reproducibility for other slippage tests, such as for the alternatives that two or more groups have slipped either to the right or to the left.

The time saved and the results obtained by the employment of the NPI-RP-B and NPI-RP-SO approaches demonstrate that these approaches are sufficient to overcome the calculation challenges associated with the NPI-RP-E approach. In this thesis, using the NPI-RP-B and NPI-RP-SO approaches the maximum runtime for the R code was approximately 6 minutes, which occurred with $k = 5$ groups and a sample size of 20 and when the number of orderings sampled is 150000 and the sample size is 30.

This research has thrown up some interesting questions that need further investigation. For example, if the reproducibility of obtaining the same result as the original experiment is low, what actions should be taken? Future analysis should be carried out, such as repeating the experiment or considering other measures if the research results continue to reveal poor reproducibility.

In practical situations, statistical tests are used to make decisions. Thus, when the RP is low, it is important to raise awareness that the results cannot necessarily be trusted as the actual evidence from the data was not very strong. One suggestion that also can be considered is to investigate the RP for sequential tests to gain a better understanding of how the actual data sample size and the future data sample size affects the reproducibility. This can be done by conducting the original test with a starting sample size and then sequentially adding observations to the actual data set. In this approach, the future data sample size will be the same as the original sample size and will consider increasing the initial data set.

In this thesis, we consider RP for different statistical tests using the NPI framework. Other perspectives can be considered as future work such as studying the RP for these tests from a Bayesian predictive perspective, which was introduced by [17]. There are also other ways to consider the RP including investigating the relationship between the RP estimates and the true RP. This could be done by simulations, drawing repeated samples from the same model and applying the test to each sample, to estimate the proportion of times when the same outcome is obtained in the test.

Appendix A

Additional Materials for Chapter 2

We introduce additional tables and figures to Section 2.5. The NPI approach to study reproducibility for the Kruskal-Wallis test and the ANOVA test, introduced in Section 2.3, is considered for different scenarios as presented in Section 2.5. In this section, data were generated under H_1 for $k = 5$ groups and sample sizes $n = 6, 10, 20$, from Normal distribution with different means $\mu_x = \mu_y = \mu_z = \mu_v = 0$, $\mu_w = 1.5$, and standard deviation 1. The null hypothesis is $H_0 : \mu_x = \mu_y = \mu_z = \mu_v = \mu_w$ against H_1 : at least one μ_i is different, $i \in \{x, y, z, v, w\}$. The level of significance is $\alpha = 0.05$. In Tables A.1, A.2 and A.3, the reproducibility probability estimates have been reported for 10 simulated data sets. Algorithm 2 is applied with $B = 1000$ and $T = 100$. The relationship between NPI-RP-B and the p-value for the KW test and the ANOVA test is examined, as displayed in Figure A.1. Increasing the size of samples from $n = 6$ to $n = 20$ leads to increasing the power of the test and more cases rejecting the null hypothesis. Increasing the number of groups and the sample size leads to a tendency for RP estimates to be higher in cases of rejection than in non-rejection, whereas RP estimates seem to be lower in cases of non-rejection than rejection, as shown in Figure A.1. The values of RP tend to increase with increasing distance between the observed p-value and the test threshold 0.05, regardless of the decision about H_0.

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
14.598	0.006	R	0.503	0.785	0.825	0.826	0.850	5.602	0.002	R	0.473	0.819	0.847	0.847	0.874
13.944	0.007	R	0.481	0.881	0.909	0.910	0.927	7.430	4.355×10^{-4}	R	0.543	0.906	0.926	0.926	0.951
10.753	0.029	R	0.371	0.733	0.780	0.782	0.809	4.974	0.004	R	0.443	0.780	0.817	0.819	0.848
9.682	0.046	R	9.682	0.612	0.656	0.657	0.691	3.443	0.023	R	0.355	0.651	0.690	0.690	0.733
9.398	0.052	NR	0.324	0.348	0.390	0.389	0.428	3.763	0.016	R	0.376	0.639	0.682	0.683	0.724
8.254	0.083	NR	0.285	0.400	0.432	0.431	0.470	2.291	0.088	NR	0.268	0.391	0.427	0.426	0.464
7.686	0.104	NR	0.265	0.391	0.432	0.432	0.465	2.418	0.075	NR	0.279	0.355	0.410	0.411	0.440
6.761	0.149	NR	0.233	0.508	0.540	0.539	0.575	1.763	0.168	NR	0.220	0.479	0.518	0.518	0.556
5.656	0.226	NR	0.195	0.511	0.543	0.544	0.580	2.130	0.107	NR	0.254	0.430	0.468	0.467	0.501
4.693	0.320	NR	0.162	0.533	0.567	0.567	0.600	1.206	0.333	NR	0.162	0.525	0.556	0.557	0.589

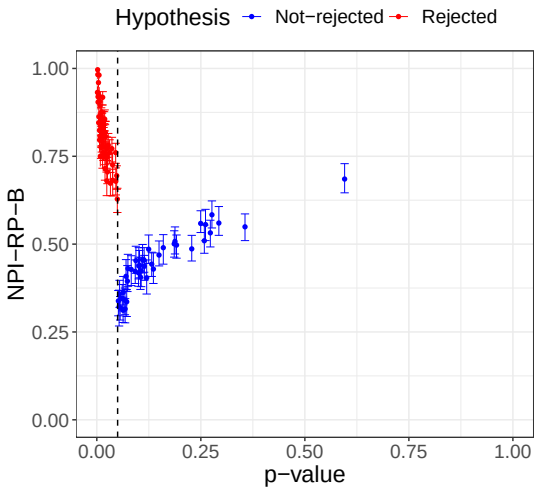
Table A.1: RP under H_1, with $k = 5$ and $X \sim N(0,1)$, $Y \sim N(0,1)$, $Z \sim N(0,1)$, $V \sim N(0,1)$, $W \sim N(1.5,1)$, $n = 6$, $\chi^2_{4,0.05} = 9.49$, $F(0.05,4,25) = 2.759$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
21.946	2.054×10^{-4}	R	0.448	0.990	0.995	0.996	0.999	6.722	2.490×10^{-4}	R	0.374	0.945	0.964	0.965	0.983
19.433	0.001	R	0.397	0.964	0.977	0.977	0.990	6.736	2.450×10^{-4}	R	0.375	0.947	0.963	0.963	0.979
17.339	0.002	R	0.354	0.877	0.907	0.908	0.933	7.863	6.794×10^{-5}	R	0.411	0.917	0.933	0.933	0.949
15.149	0.004	R	0.309	0.813	0.843	0.844	0.872	6.945	1.921×10^{-4}	R	0.382	0.826	0.868	0.870	0.895
13.784	0.008	R	0.281	0.869	0.889	0.889	0.916	5.429	0.001	R	0.326	0.892	0.910	0.910	0.931
11.826	0.019	R	0.241	0.749	0.790	0.790	0.822	4.218	0.006	R	0.273	0.776	0.811	0.811	0.836
11.607	0.021	R	0.237	0.777	0.806	0.807	0.833	3.413	0.016	R	0.233	0.741	0.770	0.770	0.796
10.416	0.034	R	0.213	0.702	0.729	0.728	0.769	3.153	0.023	R	0.219	0.679	0.722	0.722	0.749
9.563	0.048	R	0.195	0.748	0.781	0.781	0.831	2.501	0.056	NR	0.182	0.239	0.272	0.273	0.303
8.042	0.090	NR	0.164	0.329	0.378	0.379	0.420	3.055	0.026	R	0.214	0.625	0.660	0.661	0.701

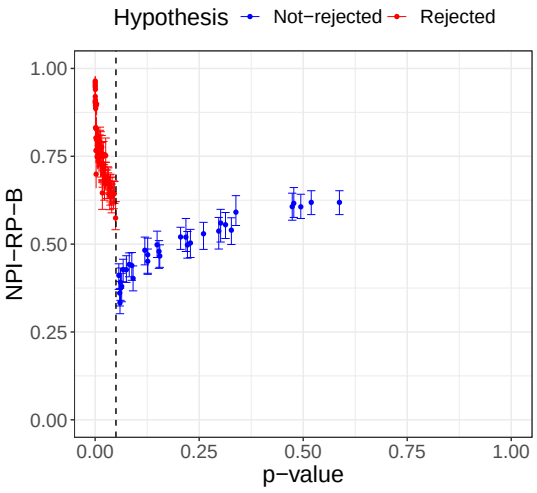
Table A.2: RP under H_1, with $k = 5$ and $X \sim N(0,1)$, $Y \sim N(0,1)$, $Z \sim N(0,1)$, $V \sim N(0,1)$, $W \sim N(1.5,1)$, $n = 10$, $\chi^2_{4,0.05} = 9.49$, $F(0.05,4,45) = 2.579$

KW test								ANOVA test							
KW	p-value	H_0	ε^2	Min	Mean	Median	Max	F	p-value	H_0	η^2	Min	Mean	Median	Max
34.329	6.379×10^{-7}	R	0.347	0.997	0.999	1	1	14.997	1.549×10^{-9}	R	0.387	0.997	1.000	1	1
32.590	1.449×10^{-6}	R	0.329	0.991	0.996	0.997	1	12.779	2.316×10^{-8}	R	0.350	0.990	0.996	0.996	0.999
29.032	7.700×10^{-6}	R	0.293	0.964	0.978	0.978	0.987	10.601	3.816×10^{-6}	R	0.309	0.962	0.974	0.975	0.988
28.461	1.006×10^{-5}	R	0.287	0.981	0.989	0.990	0.997	11.084	2.023×10^{-7}	R	0.318	0.984	0.992	0.992	0.998
26.035	3.113×10^{-5}	R	0.263	0.985	0.991	0.991	0.999	10.565	4.001×10^{-7}	R	0.308	0.986	0.992	0.993	0.998
24.317	6.901×10^{-5}	R	0.246	0.974	0.983	0.983	0.992	7.808	1.745×10^{-5}	R	0.247	0.964	0.975	0.974	0.988
21.984	2.019×10^{-4}	R	0.222	0.954	0.966	0.965	0.977	7.445	2.925×10^{-5}	R	0.239	0.954	0.964	0.964	0.977
19.970	5.061×10^{-4}	R	0.202	0.922	0.940	0.939	0.959	7.311	3.544×10^{-5}	R	0.235	0.947	0.959	0.959	0.974
15.458	0.004	R	0.156	0.868	0.886	0.886	0.910	7.445	2.925×10^{-5}	R	0.239	0.954	0.964	0.964	0.977
12.698	0.013	R	0.128	0.815	0.843	0.842	0.874	4.458	0.002	R	0.158	0.840	0.869	0.869	0.896

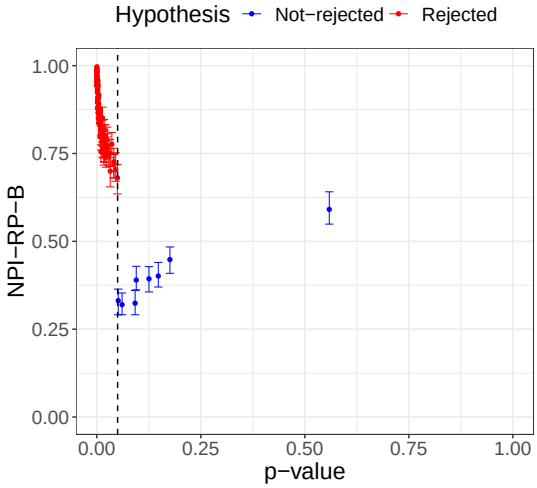
Table A.3: RP under H_1, with $k = 5$ and $X \sim N(0,1)$, $Y \sim N(0,1)$, $Z \sim N(0,1)$, $V \sim N(0,1)$, $W \sim N(1.5,1)$, $n = 20$, $\chi^2_{4,0.05} = 9.49$, $F(0.05,4,95) = 2.467$



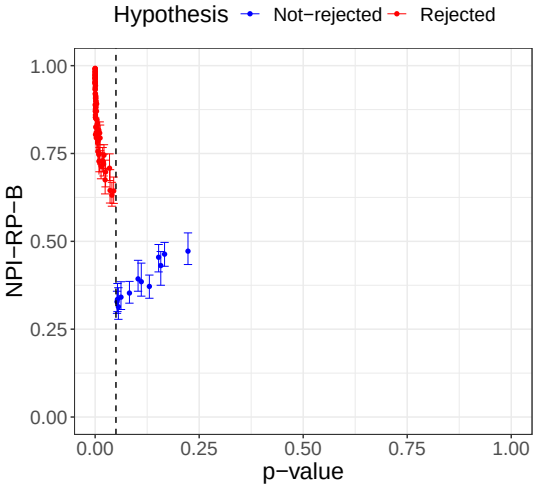
(a) KW, $n = 6$



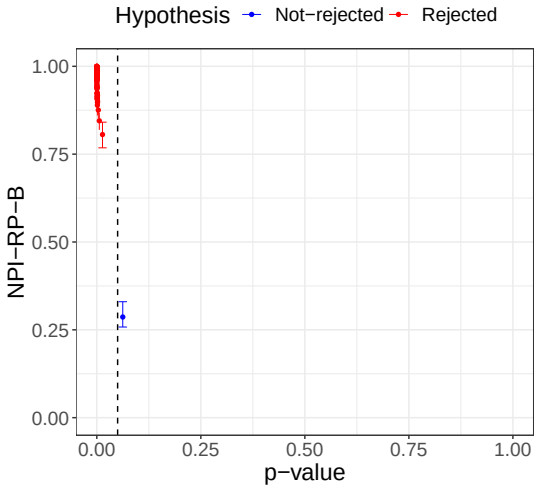
(b) ANOVA, $n = 6$



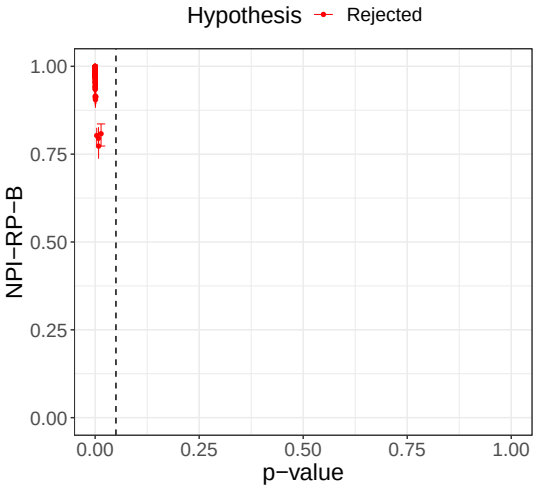
(c) KW, $n = 10$



(d) ANOVA, $n = 10$



(e) KW, $n = 20$



(f) ANOVA, $n = 20$

Figure A.1: NPI-RP-B under H_1, with $k = 5$ and $X \sim N(0,1)$, $Y \sim N(0,1)$, $Z \sim N(0,1)$, $V \sim N(0,1)$ and $W \sim N(1.5,1)$, $\alpha = 0.05$

Appendix B

Additional Materials for Chapter 3

Additional tables and figures to Section 3.5 are provided here. The NPI approach to study reproducibility for the JT test, introduced in Section 3.4, is considered for different scenarios as presented in Section 3.5. The null hypothesis is $H_0 : \mu_x = \mu_y = \mu_z$ and the alternative hypothesis is H_1: $\mu_x \leq \mu_y \leq \mu_z$, the level of significance is $\alpha = 0.05$. Simulations were performed under H_0 and H_1 for $k = 3$ groups and unequal sample sizes, with $n_x = 4$, $n_y = 8$ and $n_z = 10$.

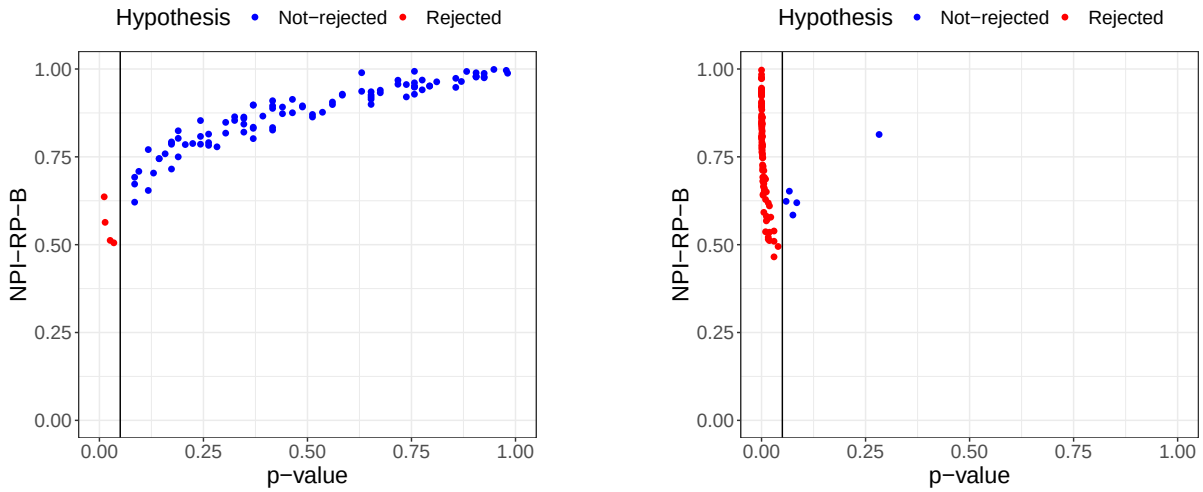
Under H_0, original data were generated from the standard Normal distribution. Under H_1, original data were generated from the Normal distribution with different means $\mu_x = 0$, $\mu_y = 1$ and $\mu_z = 2$, and standard deviation 1. Algorithm 2 is applied with $B = 1000$ and $T = 100$. The reproducibility probability estimates have been reported for 20 simulated data sets for each scenario, in Tables B.1 and B.2. The relationship between NPI-RP-B and the p-value for the JT test is examined for 100 simulated data sets, as in Figure B.1. Under H_1, we expect to observe a noticeable number of cases where the JT test rejects the null hypothesis based on the setup described earlier, and the power of the JT test is 0.882. The reproducibility probability becomes close to 0.5 in both cases of rejection and non-rejection when the observed p-value is very close to the threshold $\alpha = 0.05$, substantially below 0.5 in cases of rejection more than non-rejection. Reproducibility tends to increase when the p-value moves away from $\alpha = 0.05$, which means that the test is reproducible.

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
105	0.040	R	0.425	0.464	0.466	0.507	79	0.440	NR	0.828	0.864	0.865	0.888
102	0.059	NR	0.571	0.603	0.602	0.648	73	0.584	NR	0.896	0.919	0.919	0.939
101	0.067	NR	0.617	0.650	0.648	0.690	65	0.757	NR	0.937	0.959	0.959	0.970
99	0.085	NR	0.596	0.629	0.631	0.664	60	0.842	NR	0.924	0.948	0.948	0.966
98	0.095	NR	0.618	0.658	0.656	0.697	54	0.915	NR	0.974	0.988	0.988	0.995
95	0.130	NR	0.709	0.737	0.738	0.773	54	0.915	NR	0.978	0.988	0.988	0.995
91	0.189	NR	0.767	0.799	0.799	0.827	50	0.948	NR	0.974	0.986	0.986	0.996
91	0.189	NR	0.724	0.771	0.771	0.810	47	0.965	NR	0.985	0.993	0.993	0.999
88	0.243	NR	0.745	0.782	0.783	0.815	44	0.978	NR	0.977	0.990	0.991	0.998
85	0.303	NR	0.811	0.854	0.853	0.884	39	0.990	NR	0.993	0.997	0.998	1

Table B.1: RP for the JT test under H_0, with $k = 3$ and the original samples from $N(0, 1)$, $n_x = 4$, $n_y = 8$, $n_z = 10$, $J_{0.0456} = 104$

J	p-value	H_0	Min	Mean	Median	Max	J	p-value	H_0	Min	Mean	Median	Max
141	5.230×10^{-6}	R	0.971	0.981	0.981	0.990	120	0.003	R	0.656	0.695	0.695	0.723
141	5.230×10^{-6}	R	0.952	0.967	0.968	0.980	117	0.006	R	0.587	0.619	0.618	0.662
139	1.209×10^{-5}	R	0.960	0.972	0.973	0.985	115	0.008	R	0.540	0.575	0.575	0.608
138	1.791×10^{-5}	R	0.918	0.934	0.933	0.955	113	0.012	R	0.509	0.544	0.544	0.575
132	1.401×10^{-4}	R	0.887	0.915	0.915	0.931	110	0.019	R	0.422	0.456	0.456	0.488
129	3.367×10^{-4}	R	0.820	0.844	0.845	0.871	108	0.026	R	0.488	0.529	0.530	0.557
127	5.774×10^{-4}	R	0.753	0.788	0.789	0.816	107	0.030	R	0.453	0.495	0.495	0.537
124	0.001	R	0.801	0.829	0.830	0.862	106	0.035	R	0.426	0.463	0.463	0.499
124	0.001	R	0.745	0.773	0.775	0.795	99	0.085	NR	0.609	0.643	0.642	0.682
121	0.002	R	0.691	0.729	0.729	0.761	90	0.206	NR	0.719	0.751	0.750	0.794

Table B.2: RP for the JT test under H_1, with $k = 3$ and $X \sim N(0, 1)$, $Y \sim N(1, 1)$ and $Z \sim N(2, 1)$, $n_x = 4$, $n_y = 8$, $n_z = 10$, $J_{0.0456} = 104$



(a) under H_0

(b) under H_1

Figure B.1: NPI-RP-B for the JT test, with $k = 3$, $n_x = 4$, $n_y = 8$ and $n_z = 10$, $\alpha = 0.05$

Appendix C

Additional Materials for Chapter 4

Here, we provide the proof of Equations 4.18 and 4.19, introduced in Section 4.3. The detailed justification for these equations is as follows. To test the null hypothesis $H_0 = \mu_x = \mu_y = \mu_z$ against the alternative $H_1 : \mu_x \leq \mu_y \geq \mu_z$, that is $p = 2$ where this refers to the second group Y. In this case, the Mack-Wolfe test for three groups X, Y and Z is the sum of Mann-Whitney counts U_{XY} and U_{ZY}

$$A_p = U_{XY} + U_{ZY} = \left[R_{XY} - \frac{n_y(n_y+1)}{2}\right] + \left[R_{ZY} - \frac{n_y(n_y+1)}{2}\right]$$

where R_{XY} is the sum of the ranks of group Y when X and Y are combined, and R_{ZY} is the sum of the ranks of group Y when Y and Z are combined.

For each combination of orderings O_ℓ, the corresponding Mack-Wolfe test statistic, given in Equation 4.3, denoted by $A_{p\ell}$. In the NPI approach, there is no assumptions about the exact location of the future observations within the groups intervals (x_{j-1}, x_j), (y_{i-1}, y_i) and (z_{k-1}, z_k). However, we do have knowledge only about the number of observations within each interval. Thus, we cannot calculate a precise value of $A_{p\ell}$ related to a specific combination of orderings, but we can derive the minimum and maximum possible values; these are denoted by $\underline{A_{p_\ell}}$ and $\overline{A_{p\ell}}$, respectively. For a particular combination of orderings, the ℓ is omitted from the right hand side for simplicity of notation.

To derive the minimum value of $A_{p\ell}$ for a particular ordering, denoted by $\underline{A_{p_\ell}}$, all S_j^X future X observations in the interval (x_{j-1}, x_j), $j = 1, ..., n_x + 1$ are put at x_j, all S_i^Y future Y observations in the interval (y_{i-1}, y_i), $i = 1, ..., n_y + 1$ are put at y_{i-1}, and all S_k^Z future Z observations in the interval (z_{k-1}, z_k), $k = 1, ..., n_z + 1$ are put at z_k, as illustrated in Figure C.1. The ranks of the S_i^Y future Y observations at y_{i-1}, $i = 1, ..., n_y + 1$, when X and Y are

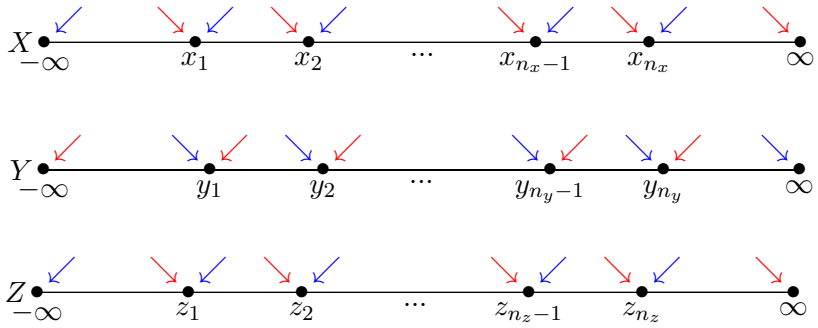


Figure C.1: The probability masses assignments for the NPI lower and upper probabilities for the events $X < Y$ and $Z < Y$

combined, are

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + 1, ..., \sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + S_i^Y \tag{C.1}$$

and ranks of the S_i^Y future Y observations at y_{i-1}, $i = 1, ..., n_y+1$, when Z and Y are combined, are

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z + 1, ..., \sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z + S_i^Y \tag{C.2}$$

then (C.1) and (C.2) sum up to

$$\begin{aligned} \underline{A_{p_\ell}} = &\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right] + \\ &\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right] \end{aligned} \tag{C.3}$$

Summing for all $i = 1..., n_y + 1$ and using the fact that $\sum_{i=1}^{n_y+1} S_i^Y = n_y$ leads to

$$\underline{A_{p_\ell}} = \sum_{i=1}^{n_y+1} S_i^Y \left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i-1)}-1} S_a^X + \sum_{c=1}^{k_{(i-1)}-1} S_c^Z\right] \tag{C.4}$$

To derive the maximum value of A_{p_ℓ} for a particular ordering O_ℓ, denoted by $\overline{A_{p_\ell}}$, all S_j^X future X observations in the interval (x_{j-1}, x_j), $j = 1, ..., n_x + 1$ are put at x_{j-1}, all S_i^Y future Y observations in the interval (y_{i-1}, y_i), $i = 1, ..., n_y + 1$ are put at y_i, and all S_k^Z future Z observations in the interval (z_{k-1}, z_k), $k = 1, ..., n_z + 1$ are put at z_{k-1}, as illustrated in Figure C.1. The ranks of the S_i^Y future Y observations at y_i, $i = 1, ..., n_y + 1$, when X and Y are combined, are

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + 1, ..., \sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + S_i^Y \tag{C.5}$$

and the ranks of the S_i^Y future Y observations at y_i, $i = 1, ..., n_y + 1$, when Z and Y are combined, is

$$\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i)}-1} S_c^Z + 1, ..., \sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i)}-1} S_c^Z + S_i^Y \tag{C.6}$$

then (C.5) and (C.6) sum up to

$$\begin{aligned}\overline{A_{p\ell}} = &\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right] + \\ &\left[\left(S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y + \sum_{c=1}^{k_{(i)}-1} S_c^Z\right] + \frac{S_i^Y(S_i^Y+1)}{2}\right) - \frac{n_y(n_y+1)}{2}\right]\end{aligned} \tag{C.7}$$

Summing for all $i = 1..., n_y + 1$ and using the fact that $\sum_{i=1}^{n_y+1} S_i^Y = n_y$ leads to

$$\overline{A_{p\ell}} = \sum_{i=1}^{n_y+1} S_i^Y\left[\sum_{b=1}^{i-1} S_b^Y - \sum_{b=i+1}^{n_y+1} S_b^Y + \sum_{a=1}^{j_{(i)}-1} S_a^X + \sum_{c=1}^{k_{(i)}-1} S_c^Z\right] \tag{C.8}$$

Appendix D

Additional Materials for Chapter 5

D.1 NPI-RP-SO for the Mosteller test

In this section, additional example is considered to study the NPI reproducibility for the Mosteller test using the NPI-RP-SO approach, introduced in Section 5.6. The NPI-RP-B method is also applied, and the results are compared to those of the NPI-RP-SO method. Data set from the literature is used for this application example, which are given in Table D.1 [99]. The null hypothesis that all groups are identical is tested against the alternative hypothesis that one group has slipped to the left, at $\alpha = 0.05$.

In Table D.2, NPI reproducibility is explored for $k = 3$ groups, using the data in Table D.1, for the average liver weights per bird for chicks given three levels of growth promoter (none, low, high), each group of size 8 [99]. For the data given in Table D.1, we break tied observations by adding a small amount [57]. Applying the Mosteller test, the group that did not receive a growth promoter is the group that has slipped to the left, with $r = 5$. The probability of obtaining 5 observations in this group which are less than all observations in the other groups, using Equation (5.1), is $P_{(r)} = 0.004$, which is less than $\alpha = 0.05$. So, the null hypothesis is rejected. In the NPI approach, there are $\binom{16}{8}\binom{16}{8}\binom{16}{8} = 2.132 \times 10^{12}$ orderings combinations to consider in the application of the NPI-RP-E. Therefore, the calculation of $\underline{RP}$ and $\overline{RP}$ becomes computationally expensive.

The NPI-RP-SO method is applied with for different number of orderings sampled r^*, and the approximation values of NPI-RP along with their confidence intervals, are presented in Tables D.2. The NPI-RP-B method has been applied using Approach I, to find the reproducibility of rejection in the future samples with the same group has slipped as the original test. Algorithm 2 is implemented with $B = 1000$ and $T = 100$. The minimum, mean, median, maximum, of $RP_1, RP_2, \ldots, RP_{100}$ are 0.415, 0.457, 0.456, 0.511, respectively. It can be concluded that accu-

None (X)	3.75	3.78	3.84	3.84	3.88	3.93	3.93	3.98
Low dose (Y)	3.92	3.96	3.96	3.99	4.02	4.03	4.06	4.10
High dose (Z)	3.94	3.94	4.02	4.06	4.08	4.09	4.12	4.17

Table D.1: Liver weights for chicks

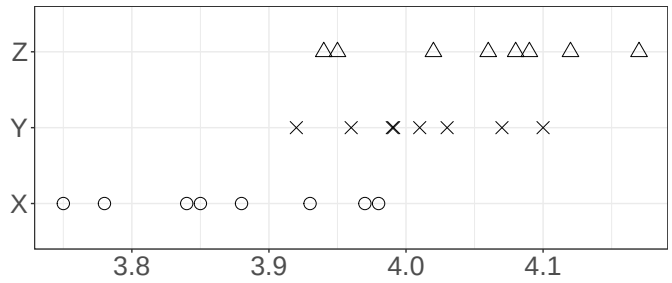


Figure D.1: Visualization of the lever weight for chicks data

rate approximations for the NPI lower and upper reproducibility probabilities for the Mosteller test can be obtained when the number of orderings sampled r^* is equal or greater than 10,000. The NPI-RP-B estimate is between the lower and upper reproducibility probabilities derived using the NPI-RP-SO method.

r^*	$\widehat{\underline{RP}}$	CI(95%)	$\widehat{\overline{RP}}$	CI(95%)
100	0.150	(0.080, 0.220)	0.900	(0.841, 0.959)
500	0.194	(0.159, 0.229)	0.926	(0.903, 0.949)
1,000	0.191	(0.167, 0.215)	0.903	(0.219, 0.272)
5,000	0.192	(0.181, 0.203)	0.899	(0.229, 0.252)
10,000	0.195	(0.187, 0.203)	0.899	(0.893, 0.905)
50,000	0.191	(0.188, 0.194)	0.896	(0.894, 0.899)
100,000	0.193	(0.191, 0.195)	0.896	(0.894, 0.898)
150,000	0.191	(0.189, 0.193)	0.896	(0.895, 0.898)

Table D.2: NPI-RP-SO for the Mosteller test with $k = 3$ and $n = 8$

D.2 NPI-RP-B for the Mosteller test

This section provides additional tables and figures to Section 5.9. The NPI approach to study reproducibility for the Mosteller test, introduced in Section 5.7, is considered for different scenarios as presented in Section 5.9. The null hypothesis is that all groups are equal and the alternative hypothesis is that one group has slipped to the left. The level of significance is

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.868	0.890	0.890	0.912
2	1	1	NR	0.856	0.882	0.882	0.912
3	1	1	NR	0.838	0.871	0.872	0.902
4	1	1	NR	0.818	0.853	0.853	0.883
5	1	1	NR	0.811	0.851	0.851	0.880
6	2	0.313	NR	0.795	0.833	0.835	0.864
7	2	0.313	NR	0.747	0.783	0.783	0.814
8	2	0.313	NR	0.750	0.782	0.782	0.817
9	3	0.092	NR	0.601	0.634	0.633	0.684
10	3	0.092	NR	0.585	0.625	0.622	0.668

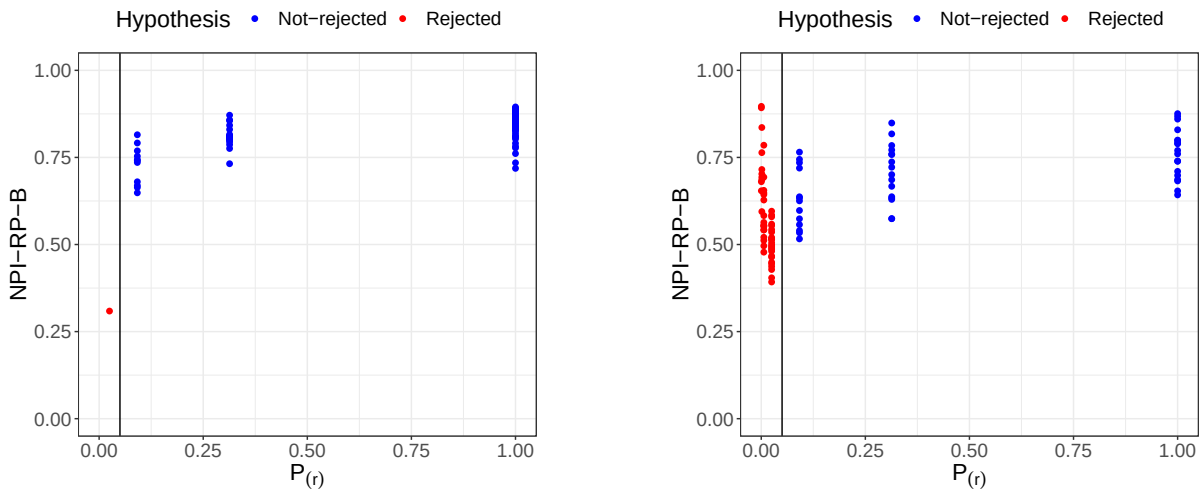
Table D.3: RP for the Mosteller test under H_0, the original samples from $N(0,1)$, $n_x = 7$, $n_y = 8$ and $n_z = 10$, $\alpha = 0.05$

$\alpha = 0.05$. Data were generated under H_0 and H_1 for $k = 3$ groups. Under H_0, original data were generated from the standard Normal distribution. Under H_1, data were generated from Normal distribution with different means $\mu_x = 0$, $\mu_y = 1.5$ and $\mu_z = 2$, and standard deviation 1. In Tables D.3 and D.4, date with unequal sample sizes is considered under H_0 and H_1, with $n_x = 7$, $n_y = 8$ and $n_z = 10$. Tables D.5 and D.6, consider data with equal sample sizes $n = 25$ under both H_0 and H_1. The reproducibility probability estimates have been reported for 10 simulated data sets. Algorithm 2 is applied using Approach II, with $B = 1000$ and $T = 100$.

The relationship between NPI-RP-B and the probability $P_{(r)}$ for the Mosteller test is examined for 100 simulated data sets, as displayed in Figures D.2 and D.3. The level of significance $\alpha = 0.05$ is represented on the figures by a vertical line. The reproducibility tend to increase with increasing distance between the observed $P_{(r)}$ and the threshold 0.05, regardless of the decision about H_0. It is clear that, as expected, the reproducibility probability is small when $P_{(r)}$ is close to the threshold. In such cases, reproducibility tends to be lower in the case of rejection than for non-rejection.

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.797	0.830	0.831	0.853
2	2	0.313	NR	0.695	0.733	0.732	0.762
3	3	0.092	NR	0.653	0.688	0.689	0.723
4	3	0.092	NR	0.594	0.643	0.644	0.687
5	4	0.025	R	0.344	0.385	0.385	0.421
6	4	0.025	R	0.358	0.388	0.388	0.424
7	5	0.006	R	0.421	0.456	0.455	0.490
8	5	0.006	R	0.503	0.544	0.545	0.585
9	6	0.001	R	0.634	0.679	0.678	0.717
10	7	2.684×10^{-4}	R	0.610	0.654	0.653	0.693

Table D.4: RP for the Mosteller test under H_1, $X \sim N(0,1)$, $Y \sim N(1.5,1)$ and $Z \sim N(2,1)$, $n_x = 7$, $n_y = 8$ and $n_z = 10$, $\alpha = 0.05$



(a) Under H_0

(b) Under H_1

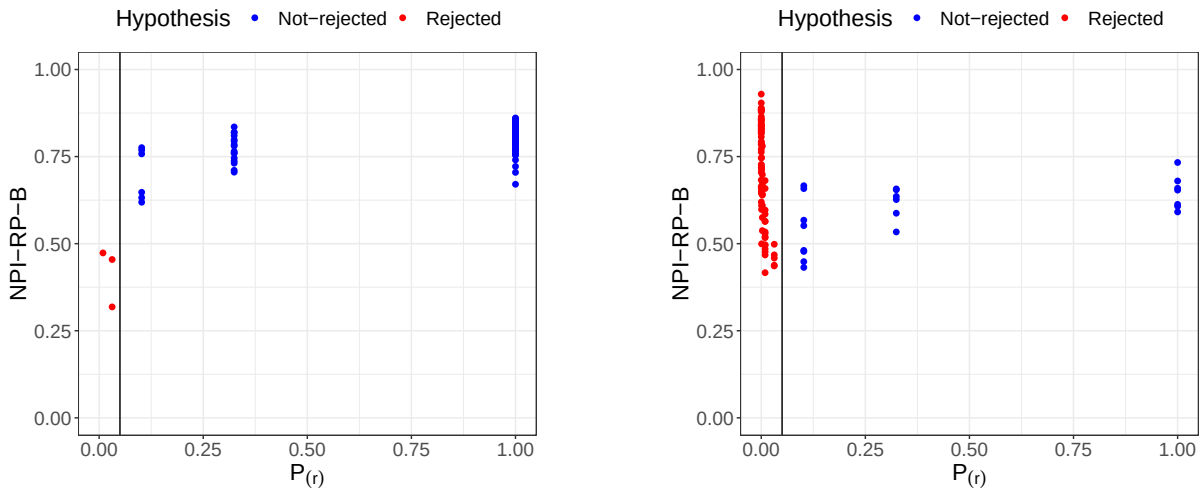
Figure D.2: NPI-RP-B for the Mosteller test, $n_x = 7$, $n_y = 8$ and $n_z = 10$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.815	0.844	0.843	0.873
2	1	1	NR	0.812	0.839	0.839	0.870
3	1	1	NR	0.807	0.832	0.832	0.865
4	2	0.324	NR	0.755	0.802	0.802	0.834
5	2	0.324	NR	0.796	0.822	0.824	0.845
6	2	0.324	NR	0.722	0.757	0.757	0.790
7	3	0.102	NR	0.697	0.727	0.725	0.755
8	3	0.102	NR	0.676	0.709	0.710	0.751
9	3	0.102	NR	0.574	0.615	0.617	0.651
10	3	0.102	NR	0.621	0.659	0.660	0.688

Table D.5: RP for the Mosteller test under H_0, the original samples from $N(0,1)$, $n = 25$, $\alpha = 0.05$

Samples	r	$P_{(r)}$	H_0	Min	Mean	Median	Max
1	1	1	NR	0.748	0.776	0.775	0.807
2	3	0.102	NR	0.556	0.595	0.598	0.640
3	4	0.031	R	0.481	0.529	0.531	0.567
4	7	0.001	R	0.611	0.658	0.657	0.697
5	8	1.923×10^{-4}	R	0.573	0.625	0.624	0.660
6	8	1.923×10^{-4}	R	0.638	0.685	0.685	0.722
7	10	1.183×10^{-5}	R	0.791	0.816	0.815	0.843
8	11	2.730×10^{-6}	R	0.808	0.835	0.834	0.859
9	13	1.232×10^{-7}	R	0.884	0.901	0.901	0.926
10	15	4.301×10^{-9}	R	0.876	0.905	0.905	0.923

Table D.6: RP for the Mosteller test under H_1, $X \sim N(0,1)$, $Y \sim N(1.5,1)$ and $Z \sim N(2,1)$, $n = 25$, $\alpha = 0.05$



(a) Under H_0

(b) Under H_1

Figure D.3: NPI-RP-B for the Mosteller test, $n = 25$, $\alpha = 0.05$

Bibliography

[1] Alabdulhadi, M. (2018). *Nonparametric predictive inference for diagnostic test thresholds*. PhD thesis, Durham University. https://npi-statistics.com/pdfs/theses/MA18.pdf.

[2] Aldawsari, A. (2023). *Parametric predictive bootstrap and test reproducibility*. PhD thesis, Durham University. http://etheses.dur.ac.uk/14970/.

[3] Alghamdi, F. M. (2022). *Reproducibility of statistical inference based on randomised response data*. PhD thesis, Durham University. http://etheses.dur.ac.uk/14783/.

[4] Alqifari, H. N. (2017). *Nonparametric predictive inference for future order statistics*. PhD thesis, Durham University. https://npi-statistics.com/pdfs/theses/HA17.pdf.

[5] Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics*, 11:375–386.

[6] Atmanspacher, H. and Maasen, S. (2016). *Reproducibility: principles, problems, practices, and prospects*. Wiley, Hoboken.

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

[8] Baker, M. (2015). Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help. *Nature*, 3.

[9] Baker, R. M., Coolen-Maturi, T. and Coolen, F. P. A. (2017). Nonparametric predictive inference for stock returns. *Journal of Applied Statistics*, 44:1333–1349.

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

[11] Bartholomew, D. (1961). A test of homogeneity of means under restricted alternatives. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 23:239–272.

[12] Bartholomew, D. J. (1959). A test of homogeneity for ordered alternatives. *Biometrika*, 46:36–48.

[13] Basso, D. and Salmaso, L. (2011). A permutation test for umbrella alternatives. *Statistics and Computing*, 21:45–54.

[14] Bergman, R. G. and Danheiser, R. L. (2016). Reproducibility in chemical research. *Angewandte Chemie International Edition*, 55:12548–12549.

[15] Bewick, V., Cheek, L. and Ball, J. (2004). Statistics review 10: Further nonparametric methods. *Critical Care*, 8:196–199.

[16] Bhat, S. V. (2009). Simple k-sample rank tests for umbrella alternatives. *Research Journal of Mathematics and Statistics*, 1:27–29.

[17] Billheimer, D. (2019). Predictive inference and scientific reproducibility. *The American Statistician*, 73:291–295.

[18] BinHimd, S. (2014). *Nonparametric predictive methods for bootstrap and test reproducibility*. PhD thesis, Durham University. https://npi-statistics.com/pdfs/theses/SB14.pdf.

[19] Bofinger, V. J. (1965). The k-sample slippage problem. *Australian Journal of Statistics*, 7:20–31.

[20] Bonnini, S., Corain, L., Marozzi, M. and Salmaso, L. (2014). *Nonparametric hypothesis testing: rank and permutation methods with applications in R*. John Wiley & Sons, West Sussex.

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

[22] Chen, Y. and Wolfe, D. (1990). A study of distribution-free tests for umbrella alternatives. *Biometrical Journal*, 32:47–57.

[23] Cohen, B. H. (2013). *Explaining psychological statistics*. John Wiley, New York.

[24] Cohen, J. (1992). Quantitative methods in psychology: A power primer. *Psychological bulletin*, 112:1155–1159.

[25] Conover, W. J. (1968). Two k-sample slippage tests. *Journal of the American Statistical Association*, 63:614–626.

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

[27] Coolen, F. P. A. (2004). On the use of imprecise probabilities in reliability. *Quality and Reliability Engineering International*, 20:193–202.

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

[29] 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.*, 16:167–185.

[30] Coolen, F. P. A. and Augustin, T. (2005). Learning from multinomial data: a nonparametric predictive alternative to the imprecise Dirichlet model. *In International Symposium on Imprecise Probability: Theories and Applications*, 5:125–134.

[31] Coolen, F. P. A. and Augustin, T. (2009). A nonparametric predictive alternative to the imprecise dirichlet model: the case of a known number of categories. *International Journal of Approximate Reasoning*, 50:217–230.

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

[33] 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.*, 14:1–13.

[34] Coolen, F. P. A. and Yan, K. (2003). Nonparametric predictive comparison of two groups of lifetime data. 3:148–161.

[35] Coolen, F. P. A. and Yan, K. J. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126:25–54.

[36] Cuzick, J. (1985). A Wilcoxon type test for trend. *Statistics in Medicine*, 4:87–90.

[37] De Capitani, L. (2013). An introduction to RP-testing. *Epidemiology Biostatistics and Public Health*, 10:1–16.

[38] 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 and Probability Letters*, 81:937–946.

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

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

[41] Doornbos, R. and Prins, H. J. (1956). *A Slippage test for a set of Gamma-variates*. Mathematisch Centrum, Amesterdam.

[42] Doornbos, R. and Prins, H. J. (1958). On slippage tests. *Indag. Math*, 20:38–55.

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

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

[45] Esra, G. and Fikri, G. (2016). A modified Mack–Wolfe test for the umbrella alternative problem. *Communications in Statistics-Theory and Methods*, 45:7226–7241.

[46] Field, Z., Miles, J. and Field, A. (2012). Discovering statistics using r. *Discovering statistics using R*, pages 1–992.

[47] Finetti, B. D. (1974). *Theory of probability: a critical introductory treatment*. Wiley, New York.

[48] Geisser, S. (1993). *Predictive inference*, volume 55. CRC press, New York.

[49] Gibb, B. C. (2014). Reproducibility. *Nature Chemistry*, 6:653–654.

[50] Gibbons, J. and Chakraborti, S. (2003). *Nonparametric statistical inference: revised and expanded. statistics: A Series of textbooks and monographs*. Marcel Dekker, Inc, New York and Basel.

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

[52] Granger, C. W. J. and Neave, H. R. (1968). A quick test for slippage. *Revue de l'Institut International de Statistique*, 36:309–312.

[53] Grant, S., C, E. and B., R. (2020). *NSM3: Functions and datasets to accompany Hollander, Wolfe, and Chicken - Nonparametric statistical methods, third edition*. R package version 1.14.

[54] Gundersen, O. E. (2021). The fundamental principles of reproducibility. *Philosophical Transactions of the Royal Society A*, 379:20200210.

[55] Hecke, T. V. (2012). Power study of anova versus kruskal-wallis test. *Journal of Statistics and Management Systems*, 15:241–247.

[56] Hettmansperger, T. and Norton, R. M. (1987). Tests for patterned alternatives in k-sample problems. *Journal of the American Statistical Association*, 82:292–299.

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

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

[59] Hollander, M., Wolfe, D. A. and Chicken, E. (2013). *Nonparametric statistical methods*. John Wiley & Sons, New Jersey.

[60] Howell, D. (1992). *Statistical methods for psychology*. Cengage Learning, California.

[61] Jonckheere, A. R. (1954). A test of significance for the relation between m rankings and k ranked categories. *British Journal of Statistical Psychology*, 7:93–100.

[62] Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16:345–353.

[63] Kim, T. K. (2017). Understanding one-way anova using conceptual figures. *Korean Journal of Anesthesiology*, 70:22–26.

[64] King, B. M., Rosopa, P. J. and Minium, E. W. (2018). *Statistical reasoning in the behavioral sciences*. John Wiley, New York.

[65] Kolassa, J. E. (2020). *An introduction to nonparametric statistics*. Chapman and Hall, New York.

[66] Kruskal, W. H. (1952). A nonparametric test for the several sample problem. *The Annals of Mathematical Statistics*, 23:525–540.

[67] Kutner, M. H., Nachtsheim, C. J., Neter, J. and Li, W. (2005). *Applied linear statistical models*. McGraw-Hill/Irwin, New York.

[68] Lau, A. (2009). What are repeatability and reproducibility. *ASTM Standardizations News*.

[69] Lecoutre, B., Lecoutre, M. P. and Poitevineau, J. (2010). Killeen's probability of replication and predictive probabilities: How to compute, use, and interpret them. *Psychological Methods*, 15:158–171.

[70] Maasen, S. and Atmanspacher, H. (2016). Social sciences: introductory remarks. *Reproducibility: Principles, Problems, Practices, and Prospects*, pages 385–389.

[71] Mack, G. A. and Wolfe, D. A. (1981). K-sample rank tests for umbrella alternatives. *Journal of the American Statistical Association*, 76:175–181.

[72] Magel, R. and Qin, L. (2003). A non-parametric test for umbrella alternatives based on ranked-set sampling. *Journal of Applied Statistics*, 30:925–937.

[73] Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60.

[74] Marques, F. J. and Coolen, F. P. A. (2020). Nonparametric predictive inference for test reproducibility by sampling future data orderings. *Journal of Statistical Theory and Practice*, 14:62.

[75] Marques, F. J., Coolen, F. P. A. and Coolen-Maturi, T. (2019). Introducing nonparametric predictive inference methods for reproducibility of likelihood ratio tests. *Journal of Statistical Theory and Practice*, 13:15.

[76] Millen, B. A. and Wolfe, D. A. (2005). A class of nonparametric tests for umbrella alternatives. *Journal of Statistical Research*, 39:1–18.

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

[78] Morisette, J. T. and Khorram, S. (1998). Exact binomial confidence interval for proportions. *Photogrammetric Engineering and Remote Sensing*, 64:281–282.

[79] Mosteller, F. (1948). A k-Sample Slippage test for an extreme population. *The Annals of Mathematical Statistics*, 19:58–65.

[80] Mosteller, F. and Tukey, J. W. (1950). Significance levels for a k-sample Slippage test. *Annals of Mathematical Statistics*, 21:120–123.

[81] Muhammad, N. (2016). *Predictive inference with copulas for bivariate data.* PhD thesis, Durham University. http://etheses.dur.ac.uk/11597/.

[82] Neave, H. R. (1972). Some quick tests for slippage. *Journal of the Royal Statistical Society Series D: The Statistician*, 21:197–208.

[83] Neave, H. R. (1973). A power study of some tests for slippage. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 22:269–280.

[84] Odeh, R. E. (1971). On Jonckheere's k-sample test against ordered alternatives. *Technometrics*, 13:912–918.

[85] Ott, R. L. and Longnecker, M. T. (2015). *An introduction to statistical methods and data analysis*. Cengage Learning, California.

[86] Page, E. B. (1963). Ordered hypotheses for multiple treatments: a significance test for linear ranks. *Journal of the American Statistical Association*, 58:216–230.

[87] Patil, P., Peng, R. D. and Leek, J. T. (2016). What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. *Perspectives on Psychological Science*, 11:539–544.

[88] Paulson, E. (1952). An optimum solution to the k-sample slippage problem for the normal distribution. *The Annals of Mathematical Statistics*, 23:610–616.

[89] Peng, R. D. (2011). Reproducible research in computational science. *Science*, 334:1226–1227.

[90] Puri, M. L. (1965). Some distribution free k-sample rank tests of homogeneity against ordered alternatives. 18:51–63.

[91] R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[92] Razali, N. M. and Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2:21–33.

[93] Sanjeev, B. and Sarmukaddam, S. (2006). *Nonparametric statistical inference*. Marcel Dekker, Inc, New York.

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

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

[96] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611.

[97] Simkus, A. (2023). *Contributions to statistical reproducibility and small-sample bootstrap*. PhD thesis, Durham University. http://etheses.dur.ac.uk/15294/.

[98] Simkus, A., Coolen, F. P. A., Coolen-Maturi, T., Karp, N. A. and Bendtsen, C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research*, 31:673–688.

[99] Sprent, P. and Smeeton, N. C. (2000). *Applied nonparametric statistical methods*. Chapman & Hall/CRC, New York.

[100] Stevens, J. R. (2017). Replicability and reproducibility in comparative psychology.

[101] Terpstra, T. (1952). The asymptotic normality and consistency of Kendall's test against trend, when ties are present in one ranking. *Indagationes Mathematicae*, 14:327–333.

[102] Thode, H. (2002). *Testing for normality*. CRC press, Boca Raton.

[103] Tibshirani, R. J. and Efron, B. (1993). *An introduction to the bootstrap*. Chapman and Hall, New York.

[104] Tomczak, M. and Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, 21:19–25.

[105] Traux, D. (1953). An optimum slippage test for the variances of k normal distributions. *The Annals of Mathematical Statistics*, pages 669–674.

[106] Tryon, P. V. and Hettmansperger, T. P. (1973). A class of nonparametric tests for homogeneity against ordered alternatives. *The Annals of Statistics*, pages 1061–1070.

[107] Walley, P. (1991). *Statistical reasoning with imprecise probabilities*. Chapman and Hall, New York.

[108] Wayne, W. D. (1990). *Applied nonparametric statistics*. PWS-KENT Publishing Company, Boston.

[109] Weichselberger, K. (2001). *Elementare Grundbegriffe einer allgemeineren Wahrscheinlichkeitsrechnung I: Intervallwahrscheinlichkeit als umfassendes Konzept.* Physika, Heidelberg.

[110] Zwaan, R. A., Etz, A., Lucas, R. E. and Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41:1–61.