Reproducibility of Hypothesis Tests for 2 × 2 Contingency Tables.

REID ALOTAIBI

Reproducibility of Hypothesis Tests for 2×2 Contingency Tables

Reid Mater B. Alotaibi

A Thesis presented for the degree of
Doctor of Philosophy

Department of Mathematical Sciences
University of Durham
England

August 2024

Dedicated to

My Father

for his prayers, support and motivation

My Mother

for all her unlimited love and prayers

Reproducibility of Hypothesis Tests for 2×2 Contingency Tables

Reid Mater B Alotaibi

Submitted for the degree of Doctor of Philosophy
August 2024

Abstract

The 2×2 contingency table is an important data structure in statistical analysis used to examine and compare the association between two binary variables. This table arranges data into four cells, which can be analysed using various statistical methods such as chi-square test, likelihood ratio test, Fisher's exact test, and McNemar test.

Nonparametric Predictive Inference (NPI) is a frequentist statistics method, which provides lower and upper probabilities for events involving one or more future observations. This thesis introduces NPI for 2×2 tables data and illustrates its use for several inference problems.

In statistics, hypothesis testing is a method of statistical inference used to draw conclusions, with the outcome being either rejecting or not rejecting the null hypothesis. Statistical reproducibility is the probability that, repeating a test under identical conditions and with the same sample size will lead to the same outcome. The reproducibility of an experiment's conclusion is a critical concept in every field of research. NPI and NPI-bootstrap methods have been used to study statistical reproducibility. In this thesis, we employ these methods to assess the reproducibility of statistical hypothesis tests based on a single 2×2 table and on multiple 2×2 tables.

Furthermore, this thesis explores the reproducibility of hypothesis tests using Bayesian inference to predict future observations through the posterior predictive distribution. By introducing NPI for 2×2 tables and employing Bayesian approaches, this thesis advances the study of statistical reproducibility for hypothesis tests. Reproducibility is low for both the NPI and Bayesian inference methods when the test statistic is near the threshold between rejecting and not rejecting the null hypothesis. On the other hand, reproducibility increases as the test statistic moves away from this threshold.

Declaration

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

Acknowledgements

Alhamdullilah, first and foremost, I am genuinely grateful to my Almighty Allah, for the countless blessings he has bestowed on me generally in my life and accomplishing this thesis.

I would like to express my sincere gratitude to my supervisors, Prof. Frank Coolen and Prof. Tahani Coolen-Maturi, for their unwavering support, invaluable guidance, kindness, patience, and mentorship throughout my PhD journey.

I extend my deepest gratitude to my parents for their constant support and prayers. I also owe thanks to my brothers and sister for their prayers and encouragement. Thank you to my wife for her support and patience throughout my PhD journey, as well as to my children. Many thanks also go to my friends who have assisted me, stood by me, and provided support and encouragement.

I am grateful to the Saudi Arabian Cultural Bureau in London and Shaqra University in Saudi Arabia for their financial support and the scholarship that allowed me to complete my studies abroad. Additionally, I want to express my thanks to Durham University for providing the facilities that enabled my smooth academic journey.

Contents

Chapter 1

Introduction

1.1 Motivation

In statistics, contingency tables are an efficient analytical tool for exploring relationships between categorical variables. It displays the association between two categorical variables, where the categories of one variable represent the rows and the categories of the other variable represent the columns [51]. The most basic version is the 2×2 table, which concerns two binary variables with two outcomes. These tables are commonly used in medical research, particularly in studying associations between binary outcomes such as the presence or absence of a disease [2]. For instance, in medical studies, 2×2 contingency tables facilitate the analysis of relationships between categorical exposure variables like smoking status and medical outcomes such as lung cancer diagnoses.

There are several common tests for analysing 2×2 contingency tables, including the chi-square test of independence, likelihood ratio test of independence, Fisher's exact test, and McNemar test. The chi-square test of independence is an important statistical tool for hypothesis testing. It compares the observed and expected frequencies within cells of categorical variables to determine whether two nominal variables are associated [1, 48]. The likelihood ratio test is similar to the chi-square test but uses the Likelihood ratio as the test statistic. Fisher's exact test is useful for small sample sizes or situations where the chi-square test is not appropriate. This test calculates the exact probabilities of observing the data under the null hypothesis [1, 48]. These tests help researchers and statisticians determine if there is a significant relationship between categorical variables or if their occurrences are independent of each other. The McNemar test is a well-known non-parametric test used to determine whether the proportions of paired nominal variables are equal. [57].

However, the choice of statistical test can influence the reproducibility of research findings. Reproducibility is a critical concept in every field of research, including the sciences and social sciences, as it is important for ensuring the accuracy and reliability of experimental findings and conclusions [6]. Recently, the reproducibility of statistical hypothesis tests has gained attention among researchers. It aasks whether different statistical tests under identical test conditions will lead to the same conclusion when repeated with respect to rejection or non-rejection of the null- hypothesis. In this thesis, we will explore the reproducibility in various tests, such as those used for 2 × 2 contingency tables. In recent years, reproducibility has been studied for various tests using nonparametric predictive inference (NPI). A detailed discussion on reproducibility is presented in Section 2.3.

Nonparametric predictive inference (NPI) is a frequentest statistics approach based on only few assumptions, enabled by the use of lower and upper probabilities to quantify uncertainty in the absence of prior knowledge. NPI, which is based on Hill's assumption $A_{(n)}$ [41, 43], has been developed to address a variety of statistical problems and data types. These include Bernoulli data [19, 20], real-valued data [20, 26], right-censored observations [22], circular data [20], multinomial data [22], and applications employing NPI-based bootstrap [23]. NPI offers an attractive framework for decision support across a wide range of problems where the focus is naturally on one or more future observations. Further details on NPI are discussed in Section 2.4

Unlike NPI, Bayesian inference provides a different approach to handling uncertainty by incorporating prior information. Bayesian inference is a method of statistical inference that uses prior probability distributions to represent and quantify uncertainty in all forms [36]. These capture initial beliefs about parameters before observing any data, which are then updated with observed data using Bayes' theorem to generate posterior distributions [36, 37, 44]. Additionally, Bayesian inference allows for predictive modeling through the posterior predictive distribution, which combines uncertainty from parameter estimation and future observation variability.

This thesis introduces an NPI method for 2×2 table data, which derives upper and lower probabilities for events involving single and multiple future observations based on 2×2 table data. The reliability of statistical hypothesis test results depends on their reproducibility, thus making reproducibility an important factor in hypothesis testing and examining statistical inferences. NPI-based methods have been used to study reproducibility, utilising its predictive nature to formulate inferences on reproducibility probability. This thesis uses the NPI and Bayesian inference methods to study the reproducibility of statistical hypothesis tests based on a single 2×2 table and on multiple 2×2 tables. Apart from reproducibility, the Bayesian method allows future observations to be predicted through the posterior predictive distribution.

1.2 Outline of thesis

This thesis is organized as follows: Chapter 2 introduces preliminary materials from the literature relevant to this thesis: 2×2 contingency tables and statistical hypothesis tests based on 2×2 tables are presented first, followed by a brief introduction to the topic of reproducibility. Then, a concise introduction to NPI is provided, focusing on NPI for circular data, NPI Bootstrap, NPI sampling of orderings and NPI for the reproducibility probability.

Chapter 3 presents a method for 2×2 tables using NPI for circular data. The NPI derive lower and upper probabilities for event involving single and multiple future observations based on 2×2 table data. We present NPI-bootstrap for circular data as a computational method for 2×2 table data. Part of this chapter was presented at the International Conference of the ERCIM WG on Computational and Methodological Statistics in London, UK, in December 2022.

Chapter 4 introduces the NPI sampling of orderings and NPI Bootstrap methods for reproducibility of tests such as the chi-square test of independence, likelihood ratio test of independence, McNemar's test, and Fisher's exact test . Through simulation studies, we evaluate different methods for reproducibility probability (RP) across different tests to gain insights into their performance. The results in this chapter were presented at a seminar at Durham University.

Chapter 5 provides an review of literature materials to the background of Bayesian inference, specifically for 2 × 2 contingency tables. The chapter further introduces Bayesian inference method for assessing the reproducibility of tests, such as the chi-square test of independence, the likelihood ratio test of independence, McNemar's test, and Fisher's exact test. Additionally, the chapter compares the performance of Bayesian inference method NPI-B method for test reproducibility. Simulation studies are conducted to understand how Bayesian inference and NPI-B compare in terms of test reproducibility. The part of this chapter was presented at a seminar at Durham University.

Chapter 6 provides a overview of multiple 2×2 tables tests, including the Mantel-Haenszel test, the Breslow-Day test, and the Woolf test. Additionally, the chapter introduces a comparison between Bayesian inference and the NPI-B methods for assessing reproducibility of these tests. The simulation studies within the chapter aim to evaluate the performance of various methods in terms of reproducibility for different tests.

In Chapter 7, we draw conclusions and discuss related research challenges. Throughout this thesis, calculations were performed using the statistical software program R.

Chapter 2

Preliminaries

The aim of this chapter is to review some of the preliminary concepts relevant to this thesis. First, we introduce 2×2 contingency tables and statistical hypothesis tests based on 2×2 tables. After that, discuss a general review of reproducibility in statistics. Finally, we provide an overview of the methodology of nonparametric predictive inference (NPI), including NPI for circular data, NPI Bootstrap, NPI sampling of orderings and NPI for the reproducibility probability.

2.1 2×2 Contingency tables

Contingency tables, also known as cross-tabulations, are matrices that display the frequency distribution of two or more categorical variables. They are used to summarise relationships between two or more discrete variables and determine whether events are dependent or independent to measure the degree of association [35]. Karl Pearson [1] introduced a two-way contingency table in 1904, which represents the frequency of events in two dimensions, with rows representing different levels of the first variable and columns representing different levels of the second variable. The applications of contingency tables are widespread and can be found in various fields, including surveys, business analytics, medical, psychological, educational, and social sciences [49].

Howard [46] describes four types of 2×2 contingency tables: double dichotomy, where individuals in one population are classified in two ways; two binomials, which compares successes across two independent populations; comparative trials, where a single population is split into treatment groups; and tea tasting, which tests classification accuracy when treatment proportions are known. Howard [46] emphasizes that the "double dichotomy" structure is particularly distinct, as it involves categorizing individuals within a single population in two distinct ways. Different types of 2×2 tables include tables with fixed totals for rows or columns and tables with paired data, like in before-and-after studies.

Consider two categorical variables, X and Y, with $r \geq 2$ and $c \geq 2$ levels, respectively. That is the rows correspond to the categories of variable X, and the columns correspond to the categories of variable Y. Here, a contingency table with r rows and c columns is referred to as an $r \times c$ table. For instance, in the case of $r = 2$ and $c = 2$, this gives a 2×2 table, which is the focus of this thesis.

In social sciences and biomedical applications, 2×2 contingency tables are very popular for comparing two binary variables with two outcomes [48]. Table 2.1 displays general notation for the 2×2 contingency table. The cell frequency, denoted by n_{ij} where $i, j = 1, 2$, is observed counts of the number of cases with combination $(X, Y) = (i, j)$. The row margins are denoted by n_{i+} and n_{+j} for the column margins where the subscript "+" is the sum over the index it replaces. The marginal row total of the i^{th} row, $i = 1, 2$ is $n_{i+} = n_{i1} + n_{i2}$, while the marginal total of the j^{th} column is $n_{+j} = n_{1j} + n_{2j}$, $j = 1, 2$, and the total number of observations of the data set is $n = n_{1+} + n_{2+} = n_{+1} + n_{+2}$.

Table 2.2 shows the notation for the joint probabilities for the 2×2 contingency table. Let π_{ij} denote the joint probability that (X,Y) falls in the cell in row i and column j where $i, j = 1, 2$. The probability π_{i+} is the row marginal distribution, and the probability π_{+j} is the column marginal distribution, for $i, j = 1, 2$. The associated sampling proportions are denoted by $p_{ij} = \frac{n_{ij}}{n}$. The marginal proportions by p_{i+} for row i is $p_{i+} = \frac{n_{i+}}{n}$ for $i = 1, 2$ and the marginal proportions by p_{+j} for column j is $p_{+j} = \frac{n_{+j}}{n}$ for $j = 1, 2$.

For 2×2 tables, multinomial sampling is crucial because it allows accurate modeling and statistical testing of relationships between two categorical variables, especially in evaluating associations and independence. Multinomial sampling is

	c		
r	1	2	Total
1	n_{11}	n_{12}	n_{1+}
2	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Table 2.1: 2×2 contingency table

	c		
r	1	2	Total
1	π_{11}	π_{12}	π_{1+}
2	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Table 2.2: The probabilities corresponding to 2×2 contingency table.

a statistical sampling technique used to select elements from a population with multiple categories. In multinomial sampling, each component of the population is distributed to one of the various categories. Multinomial sampling is widely used in different research fields, such as medical research, market research, quality control, etc[2].

Multinomial sampling is used when the total sample size n is fixed, but the row and column totals are not [1]. We have n trials, and each trial has more than two possible outcomes. When trials are independent and each category's probability remains constant across trials, the distribution of counts across these categories follows a multinomial distribution [1, 2]. Let the probabilities be denoted by$\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$, where $\sum_{ij} \pi_{ij} = 1$. For a 2×2 contingency table, the probability mass function for the counts $n_{11}, n_{12}, n_{21}, n_{22}$, under the constraint that $\sum_{ij} n_{ij} = n$ according to the multinomial distribution, is given by:

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{n!}{n_{11}!\, n_{12}!\, n_{21}!\, n_{22}!}\, \pi_{11}^{n_{11}}\, \pi_{12}^{n_{12}}\, \pi_{21}^{n_{21}}\, \pi_{22}^{n_{22}}. \tag{2.1.1}$$

The odds ratio is a measure of association used to quantify the relationship between two categorical variables in a 2×2 table. It is commonly used to indicate the strength and direction of an association [1]. The odds ratio compares the odds of a particular event occurring in one category of a variable to the odds of the same event occurring in another category. The "odds" refer to the ratio of the probability of success to the probability of failure for a particular category and are defined as:

$$\Omega = \frac{\pi}{1-\pi},$$

where π is the probability of success, and $1 - \pi$ is the probability of failure [2].

In a 2×2 table, the odds of success are typically compared between two categories. Let π_1 represent the probability of success in category one (row one), and π_2 represent the probability of success in category two (row two). The odds of success for category one, Ω_1, are $\frac{\pi_1}{1-\pi_1}$, and the odds for category two, Ω_2, are $\frac{\pi_2}{1-\pi_2}$. The odds ratio, θ, compares these two odds:

$$\theta = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}}.$$

The odds ratio θ measures the relative odds of success in one category compared to another. If $\theta > 1$, success is more likely in the first category. If $\theta = 1$, the odds of success are the same in both categories, indicating no association between the variables, which implies independence. If $\theta < 1$, success is less likely in the first category [48]. The odds ratio is not only a numerical calculation but a measure of the strength of association between two categories.

For a 2×2 table, the odds ratio can also be derived by considering the joint probabilities of success and failure in both categories [48]. The odds ratio in this context is:

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The odds ratio is often estimated using the observed frequencies in a contingency table. Let n_{11}, n_{12}, n_{21}, and n_{22} represent the observed frequencies in the corresponding cells of the 2×2 table [48]. The sample odds ratio, $\hat{\theta}$, is calculated as:

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}.$$

This estimate assumes that the observed cell counts n_{ij} reflect the underlying probabilities. The estimated odds ratio $\hat{\theta}$ provides a measure of the strength of association between the two categories based on the sample size [48].

2.2 Hypothesis tests based on 2×2 tables

In this section, we provide an overview of 2×2 contingency table tests. The chi-square test of independence, likelihood ratio test of independence, Fisher's exact test, and McNemar test are some common methods for analyzing 2×2 tables, as described in Table 2.3. In statistical inference, hypothesis tests are used to evaluate the plausibility of a hypothesis about a population using sample data. The null hypothesis H_0 asserts that there is no effect, no difference, or no relationship among the variables. In contrast, the alternative hypothesis H_1 proposes that there is an effect, a difference, or a relationship. In terms of rejection or nonrejection of the null hypothesis, the p-value is a different approach from the critical value, but they both yield the same result about whether or not the null hypothesis is rejected. The null hypothesis must be rejected if the p-value is less than a predetermined significance level [74].

The significance level α has a predetermined value in the hypothesis testing procedure. A type I error is the probability of rejecting a true null hypothesis, and a type II error is the probability of not rejecting a false null hypothesis. The error probabilities are indicated by the symbols α and β, respectively. The Type I and Type II errors have an inverse relationship, meaning that when one increases, the other decreases. The power is defined as the probability of rejecting the false null hypothesis when an alternative hypothesis is true. In other words, the power is the probability of making the correct decision regarding the false null hypothesis, this probability is $1 - \beta$. The power of the test increases with the α level associated with the hypothesis testing procedure and with the sample size of the experiment [74].

2.2.1 Tests of independence

A test of independence assesses whether there is an association between two categorical variables X and Y [1]. It evaluates whether the distribution of one variable is independent of the other variable. Under the assumption of independence, the joint probability π_{ij} of observing category i for X and category j for Y can be expressed as the product of their marginal probabilities, π_{i+} and π_{+j}. When two categorical variables, X and Y, are statistically independent, it implies that the probability distribution of Y remains identical regardless of the category of X.

The null hypothesis H_0 states that there is no association between the categorical variables X and Y in 2×2 tables. The null hypothesis of statistical independence is [1]:

$$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}, \quad \text{for} \quad i, j \in \{1, 2\}$$

The alternative hypothesis H_1 asserts that there is an association between categorical variables X and Y. The alternative hypothesis of statistical dependence is expressed as follows:

$$H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}, \quad \text{for} \quad i, j \in \{1, 2\}$$

To test the null hypothesis H_0, the values $e_{ij} = n\pi_{ij} = n(\pi_{i+}\pi_{+j})$ is the expected frequency. Under the null hypothesis of independence, the expected cell counts are derived by multiplying the marginal totals for each row and column and then dividing by the total sample size n. This results in estimated expected cell counts calculated as $\hat{e}_{ij} = \frac{n_{i+} \times n_{+j}}{n}$ [1].

Chi-squared test of Independence

The Chi-squared independence test is essential for assessing hypotheses in 2×2 tables and is also known as a large sample independence test. It examines the independence between row and column variables by comparing the distinction between observed and expected frequencies of categorical variables [1, 48]. The test statistic for the null hypothesis H_0 is given by:

$$\chi^2 = \sum_i \sum_j \frac{(n_{ij} - \hat{e}_{ij})^2}{\hat{e}_{ij}} \qquad i, j \in \{1, 2\}$$

χ^2 follows an asymptotic chi-squared distribution with degrees of freedom $(r-1) \times (c-1)$.

Here, r and c are the numbers of rows and columns in the contingency table, respectively [1]. The degrees of freedom of a 2×2 table is 1. The test statistic χ^2 is zero if observed and expected frequencies are equal, which is the ideal scenario under the null hypothesis of independence. The larger differences between these values result in a larger test statistic, providing evidence against the null hypothesis H_0. In order to approximate the chi-square distribution, each cell should have a sufficiently large expected frequency at least 5. The chi-square independence test is a powerful tool for assessing the independence of categorical variables in 2×2 tables, with the test statistic provides information on the degree of association.

Statistical power refers to the probability that a false null hypothesis will be rejected by the test. The power depends on the effect size, sample size, and significance level. For chi-square tests, the effect size for a 2×2 table is calculated as [18]:

$$w = \sqrt{\sum_{i=1}^{2} \sum_{j=1}^{2} \frac{(p_{0i} - p_{1i})^2}{p_{0i}}}$$

where p_{0i} is the proportion in cell i given by the null hypothesis and p_{1i} is the proportion in cell i according to the alternative hypothesis. The effect size w, related to the usual chi-square statistic, is given by $w = \sqrt{\frac{\chi^2}{n}}$. Cohen [18] suggests using effect sizes of $w = 0.1$ for small effects, $w = 0.3$ for medium effects, and $w = 0.5$ for large effects. The noncentrality parameter λ, which is used to calculate the power of the chi-square test, is $\lambda = n \times w^2$ [18]. The power of the test is calculated by the probability that the test statistic exceeds the critical value of the chi-square distribution with the specified degrees of freedom, accounting for the non-centrality parameter as:

$$P\left(\chi^2_{1,\lambda} > \chi^2_{\alpha,1}\right)$$

where $\chi^2_{1,\lambda}$ is the non-central chi-square distribution with 1 degree of freedom and the non-centrality parameter λ, and $\chi^2_{\alpha,1}$ is the critical value from the chi-square distribution with 1 degree of freedom at the significance level α [2, 18].

Likelihood ratio test of independence

The likelihood ratio test assesses the independence between two categorical variables based on maximum likelihood estimation [1]. Let ℓ_0 denote the maximized value of the likelihood function under the null hypothesis, and let ℓ_1 denote the maximized value under the alternative hypothesis. The likelihood-ratio test statistic is given by:

$$-2 \log \left(\frac{\ell_0}{\ell_1} \right).$$

When the null hypothesis H_0 is not true, the maximized likelihood under the null hypothesis ℓ_0 is smaller than the maximized likelihood under the alternative hypothesis ℓ_1. This means that the ratio $\frac{\ell_0}{\ell_1}$ is less than one. Taking the logarithm of a number less than one yields a negative value: $\log \left(\frac{\ell_0}{\ell_1} \right) < 0$. Multiplying this negative value by -2 results in a positive test statistic [1]. As the sample size increases, the difference between ℓ_0 and ℓ_1 becomes more pronounced, leading to a higher positive value for the test statistic. Therefore, when H_0 is not true, the test statistic value is typically a large positive number, providing strong evidence against the null hypothesis [1].

The likelihood ratio test makes use of the fact that, under the null hypothesis of independence, the likelihood ratio statistic follows an asymptotic chi-square distribution [1]. The likelihood ratio statistic for a 2×2 contingency table can be written as:

$$G^2 = 2 \sum_{i,j} n_{ij} \log \left(\frac{n_{ij}}{\hat{e}_{ij}} \right), \quad i, j \in \{1, 2\}.$$

This statistic is called the likelihood ratio chi-squared statistic. The G^2 takes its minimum value of 0 when all observed counts n_{ij} are equal to the expected counts $\hat{e}_{ij}$, and larger values provide stronger evidence against H_0 [48]. If the sample size is small and this condition is not met, the test might not be reliable. In such cases, other tests like Fisher's Exact Test can be used to accurately determine independence.

2.2.2 Fisher's exact test

Fisher's exact test is used to determine if there is a significant association between two categorical variables. When the sample size is small, large sample tests such as χ^2 and G^2 are inconvenient due to their reliance on asymptotic properties and approximations, making exact tests more suitable. Formally, Fisher's exact test considers the hypotheses [1]:

$$H_0 : \theta = 1 \quad \text{vs.} \quad H_1 : \theta > 1$$

By conditioning on both sets of marginal totals (row and column totals) in the 2×2 table, we fix these totals and consider the distribution of cell counts given these margins. This conditioning is done because, under the null hypothesis of independence, the cell counts then follow a hypergeometric distribution, simplifying the calculation of exact probabilities [1]:

$$p(t) = P(N_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}} \tag{2.2.1}$$

For a given set of marginal totals, larger values of N_{11} correspond to larger sample odds ratios, indicating stronger evidence in favor of H_1. Therefore, the p-value is determined by the probability of observing $N_{11} \geq t_{o11}$, where t_{o11} represents the observed value of N_{11} [48]. Similarly, for $H_1 : \theta < 1$, the p-value is calculated by summing all tables where N_{11} is less than or equal to t_{o11}.

For the null hypothesis H_0 of independence, the p-value is calculated as the sum of hypergeometric probabilities for all possible contingency tables with the same marginal totals that are as or more favorable to the alternative hypothesis H_1 than the observed table. Specifically, for the case of $H_1 : \theta > 1$, where θ represents the odds ratio, tables with larger values of N_{11} (given fixed marginal totals) provide stronger evidence to support H_1. The p-value is determined as the hypergeometric probability of the right tail, evaluating whether N_{11} is at least as large as the observed value [2].

2.2.3 McNemar's Test

McNemar's test is a widely used nonparametric test to analyze paired nominal data to detect changes [57]. In a 2×2 table, pairing means having two related observations, such as measurements from the same person before and after a treatment or responses from matched subjects in a study. It assumes marginal homogeneity under the null hypothesis, which means equal marginal frequencies for rows and columns. The McNemar test is generally applied to a 2×2 table when two cells in the diagonal arrangement are considered to be two concordant responses and the other two cells are considered as discordant responses. This pairing is important because it ensures that each discordant pair (n_{12} and n_{21}) compares two related conditions or time points for the same subject. The null and alternative hypotheses are $H_0 : \pi_{12} = \pi_{21}$ and $H_1 : \pi_{12} \neq \pi_{21}$, respectively. The test statistic follows a chi-square distribution with 1 degree of freedom. Recalling the cells n_{12} and n_{21} from Table 2.1, which represent discordant pairs, the McNemar test statistic is:

$$X^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \tag{2.2.2}$$

An alternative way to calculate McNemar's test statistic is using a normal approximation:

$$T_{MN} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \tag{2.2.3}$$

The T_{MN} is asymptotically distributed as a standard normal random variable under the null hypothesis H_0 [16]. The power of McNemar's test can be approximated by:

$$\Phi\left(\frac{\sqrt{n}(\pi_{12} - \pi_{21}) - z_{\alpha/2} \cdot \sqrt{\pi_{12} + \pi_{21}}}{\sqrt{\pi_{12} + \pi_{21} - (\pi_{12} - \pi_{21})^2}}\right) \tag{2.2.4}$$

where Φ represents the cumulative distribution function of the standard normal distribution, and $\pi_{12} - \pi_{21}$ is the mean difference of the discordant pairs, with variance $\pi_{12} + \pi_{21} - (\pi_{12} - \pi_{21})^2$ [16]. The $z_{\alpha/2}$ is the critical value from the standard normal distribution.

Table 2.3: Summary of Hypothesis Tests based on 2×2 Tables

Test	**Hypothesis**	**Test Statistic**	**Assumptions**	**Properties**
Chi-squared of Independence	$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$	$\chi^2 = \sum_{i,j} \frac{(n_{ij}-\hat{e}_{ij})^2}{\hat{e}_{ij}}$	Large sample size; expected cell counts $\hat{e}_{ij} > 5$.	Compares observed and expected frequencies; follows chi-squared dist. with 1 df for 2x2 tables.
Likelihood-Ratio of Independence	$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, $H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j}$	$G^2 = 2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{e}_{ij}}\right)$	The test assumes a sufficient sample size, and expected frequencies of at least 5.	Evaluating independence using the likelihood of data under null vs. alternative hypotheses.
Fisher's Exact Test	$H_0 : \theta = 1$, $H_1 : \theta > 1$ (or $\theta < 1$)	$p(t) = P(N_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$	Small sample sizes with fixed marginals.	Exact test; Provides exact p-values and reliable for small sample sizes.
McNemar's Test	$H_0 : \pi_{12} = \pi_{21}$, $H_1 : \pi_{12} \neq \pi_{21}$	$\chi^2 = \frac{(n_{12}-n_{21})^2}{n_{12}+n_{21}}$, $T_{MN} = \frac{n_{12}-n_{21}}{\sqrt{n_{12}+n_{21}}}$	Paired nominal data with marginal homogeneity.	Nonparametric test for paired nominal data; chi-squared or normal approximation; focuses on discordant pairs.

2.3 Reproducibility

The *reproducibility* of a study refers to the capacity of the result and conclusion obtained from an experiment to be reproduced when the analysis is conducted again. Understanding reproducibility is essential for scientific research. It ensures that research findings are reliable, trustworthy and can be verified by other researchers [6]. Lack of reproducibility can lead to loss of trust in research and can affect decision-making [61]. Reproducibility is important in every field of study, whether social science, pure science, applied science, medicines, psychology or arts [12, 47, 62]. In clinical trials or drug testing studies, for instance, lack of reproducibility can be very challenging, especially if crucial decisions are to be made [47].

Begley and Ellis [9] mentioned the challenges of reproducing test results in preclinical cancer research. They analysed various factors affecting reproducibility of studies and offered recommendations for improving reproducibility. In addition, they addressed the influence of publication processes, which tend to prefer the publication of positive results, leading to bias. However, Begley and Ellis [9] only discussed general reproducibility and did not consider the statistical aspects and methodologies employed in medical testing. According to Stodden [70], statistical methods and analysis should be considered because they can affect reproducibility [70].

Statistical reproducibility is concerned with statistical concepts and methodologies, such as hypothesis testing, confidence intervals, effect of sample sizes, and p-values. Goodman [40] was the first who studied statistical reproducibility. He indicated that there was a misunderstanding about the meaning of a statistical p-value. A statistical p-value is the probability of getting the same or a more extreme value for the test statistic, under the assumption that the null hypothesis is true. Goodman, however, challenged this claim and argued that p-value overstates evidence against the null hypothesis. He mentioned that the probability that the test for an event will produce the same result if repeated under identical condition with the same sample size may be small and referred to this as replication probability [40].

Boos [64] shared the same view with Goodman and mentioned the importance of adjusting the use of p-values to account for replication in future experiments. Stahel [68] also mentioned possible misuse of p-value in expressing the results of statistical data analyses and how this can affect reproducibility [68]. Senn [65] also agreed with Goodman that reproducibility is an important aspect of test results, but disagreed with the claim that p-value overstates the evidence against the null hypothesis. Senn [65] showed a real-life scenario of reproducibility probability tests under various conditions and emphasised the significance of test reproducibility, where repeated tests may occur under slightly different conditions and involve other teams of analysts.

In addition, Miller [58] mentioned that it is necessary to differentiate between two scenarios when repeating a test: a general repetition carried out by different researchers, where conditions may vary compared to the original experiment and test, and an individual repetition conducted by the same researcher under identical conditions as the original experiment and test. However, Miller [58] expressed doubt regarding the possibility of drawing valuable inferences from a single initial investigation, mainly due to the unknown actual effect sizes and, consequently, the unknown power of the test.

Study design, statistical methods, and tests used for a study are important to consider when discussing reproducibility. This is because different methods for different studies can introduce variability, which can affect how reproducibility is measured. According to Stanley et al [69], one of the statistical factors that can affect reproducibility is low statistical power. They explained that original studies with low power could lead to biased effect size estimates, which could underestimate the sample size for replications.

Recently, there were concerns about the reproducibility of COVID-19-related research studies [60, 72]. This was because important decisions during the COVID-19 pandemic depend on the conclusions of scientific studies. Therefore, the results of experiments, such as the effectiveness of a vaccine or the impact of a public health intervention, must be reproducible when repeated using the same methods and data. However, the speed of COVID-19 research and publications and the urgency of the pandemic may affect experimental design, statistical analysis, and interpretation, making reproducibility a challenge [32]. To ensure that decisions are based on reproducible evidence, it is therefore crucial to identify reliable methods and tools to measure reproducibility.

Reproducibility measurement is challenging. According to Goodman [64], the challenge in measuring reproducibility is due to the different factors associated with the concept of reproducibility. Goodman [64] stated that statisticians play an important role in identifying common problems and solutions in various research fields. Their expertise can help to develop methods to measure and enhance the understanding of reproducibility.

Shao and Chow [66] introduced three methods for assessing reproducibility in study designs in clinical trials. These methods are the estimated power approach, also mentioned by Stanley et al. [69], the lower confidence bound of power estimate, and the Bayesian approach. According to Shao and Chow [66], a single clinical trial is enough if a statistical result from the first clinical trial is strongly reproducible. They examined the general use of clinical results from one patient population to another and changed the sample size for the second trial. De Martini [31] regarded the test's power and the lower confidence bound of the power as two definitions of the reproducibility probability for statistically significant outcomes. De Capitani and De Martini [28, 29, 30] examined different estimators of the reproducibility probability for the Wilcoxon rank sum test and explored various nonparametric tests, including the sign and Wilcoxon signed-rank tests.

In recent years, the Bayesian approach to statistical analysis has gained popularity as an effective method in addressing reproducibility [34]. One of the main challenges in statistical reproducibility is how conclusions and inferences are derived after analysing experimental data. This challenge is closely linked to the issue of p-values discussed by Goodman [40]. Statisticians employ two main approaches, namely frequentist and Bayesian methods. The frequentist approach relies on the available data or evidence to estimate parameters. It focuses on hypothesis testing, confidence intervals, and p-values. The Bayesian approach considers model parameters as random variables with associated probability distributions. It involves using prior distributions with observed data to update posterior distributions, which form the basis of Bayesian inference. Several studies have used Bayesian methods and can also be applied in machine learning, data science, and scientific modeling [52]. Höpfl and et al [45] also showed the strength and implications of Bayesian methods in examining reproducibility.

Reproducibility is naturally viewed as a problem of predictive inference. BinHimd and Coolen [11] considered reproducibility as a predictive problem and used nonparametric predictive inference (NPI), a frequentist approach, to evaluate it. In this thesis, the reproducibility probability (RP) of a hypothesis test is defined as the probability that repeating the experiment and performing the same hypothesis test yields the same conclusion, whether rejecting or not rejecting the null hypothesis [11]. The NPI approach is used to calculate the lower and upper reproducibility probabilities. To illustrate, we use the sign test, a basic non-parametric test. Consider n ordered observations $X_1, X_2, \ldots, X_n$. The test statistic W counts the number of positive observations and is given by:

$$W = \sum_{j=1}^{n} \mathbb{1}\{X_j > 0\},$$

where $\mathbb{1}(\cdot)$ is an indicator function equal to 1 if the event occurs and 0 otherwise [11]. The null hypothesis H_0 is rejected at a significance level α if $W \geq w_\alpha$, where w_α represents the upper α percentile of the binomial distribution with a sample size of n and a success probability of 1/2. The NPI approach introduced in Section 2.4 can be used to make inferences about the m future observations among the n data observations. There are $\binom{n+m}{n}$ possible orderings of m future observations among the n data observations, with all orderings equally likely, where $i = 1, 2, \ldots, \binom{n+m}{n}$ [11]. The reproducibility probability of a statistical test refers to the probability that the same test results would be obtained in a repeated test. We focus on the case where the number of future observations m is equal to the number of data observations n. The goal is to determine the minimum and maximum values of the test statistic W for each possible ordering O_i, which are denoted by $\underline{W}_i$ and $\overline{W}_i$, respectively [11].

If the original test conclusion is the rejection of H_0, then the NPI lower reproducibility probability is derived by counting the number of orderings for which $\underline{W}_i \geq w_\alpha$ [11]. The corresponding NPI upper reproducibility probability is derived by counting the number of orders for which $\overline{W_i} \geq w_\alpha$. Thus, the NPI lower and upper reproducibility probabilities are given by:

$$\underline{RP} = \frac{1}{\binom{2n}{n}} \sum_{i=1} \mathbb{1}\{\underline{W}_i \geq w_\alpha\}, \quad \overline{RP} = \frac{1}{\binom{2n}{n}} \sum_{i=1} \mathbb{1}\{\overline{W_i} \geq w_\alpha\}.$$

where $i = 1, 2, \ldots, \binom{n+m}{n}$. Similarly, if the original test conclusion is non-rejection of H_0, such that $W < w_\alpha$, the lower reproducibility probability of the NPI is derived by counting the number of orders for which $\underline{W}_i < w_\alpha$, and the corresponding upper reproducibility probability is derived by counting the number of orders for which $\overline{W_i} < w_\alpha$ [11]. The NPI lower and upper reproducibility probabilities are:

$$\underline{RP} = \frac{1}{\binom{2n}{n}} \sum_{i=1} \mathbb{1}\{\underline{W}_i < w_\alpha\}, \quad \overline{RP} = \frac{1}{\binom{2n}{n}} \sum_{i=1} \mathbb{1}\{\overline{W_i} < w_\alpha\}.$$

When using observations of n data from the first test, the NPI-RP approach calculates the lower and upper probabilities for the test statistic that rejects or rejects the null hypothesis. These probabilities are based on all possible orderings of n future observations combined with the n data observations.

2.4 Nonparametric Predictive Inference (NPI)

Nonparametric predictive inference (NPI) is a frequentist statistical method proposed to make an inference for one future observation based on past data. NPI method uses only a few assumptions to learn from the data without prior knowledge. NPI is a statistical methodology based on Hill's $A_{(n)}$ assumption [41, 42]. Hill introduced the assumption $A_{(n)}$ for predicting a single future observation X_{n+1} in the absence of prior knowledge of the underlying distribution. Let $X_1, \ldots, X_n$ be exchangeable continuous random quantities and let $x_1, \ldots, x_n$ be the corresponding observations. According to Hill's assumptions, random quantities are exchangeable, which allows learning from observed data without imposing specific dependencies or constraints. The ordered observations are denoted by $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$, and let $x_{(0)} = -\infty$ and $x_{(n+1)} = \infty$ for notation. For a future observation X_{n+1}, the assumption $A_{(n)}$ is [41, 42]

$$P(X_{n+1} \in (x_{(j-1)}, x_{(j)})) = \frac{1}{n+1} \text{ for } j = 1, 2, \ldots, n+1 \tag{2.4.1}$$

Where $(x_{(j-1)}, x_{(j)})$ can be referred to by I_j.

NPI is suitable to provide imprecise probability for an event A. Imprecise probability is a generalization of a classical probability in the meaning that it can be used to describe uncertainty about events via intervals, instead of a single probability. The lower and upper probabilities, which are considered in interval probability theory [73], can be obtained based on NPI, where NPI leads to a strong consistency property in the frequentist statistical theory [25, 73]. In NPI, the lower probability is the maximum lower bound and the upper probability is the minimum upper bound for A. The precise probability for A, $P(A)$, is a special case of imprecise probability when the lower and upper probabilities are equal. 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 the difference $\Delta(A) = \overline{P}(A) - \underline{P}(A)$ is called the imprecision [25]. Augustin and Coolen [7, 8] introduced the NPI lower and upper probabilities for the event $X_{n+1} \in B$, where $B \subseteq \mathbb{R}$, as follows :

$$\underline{P}(X_{n+1} \in B) = \frac{\sum_{j=0}^{n} \mathbb{1}(I_j \subseteq B)}{n+1} \tag{2.4.2}$$

$$\overline{P}(X_{n+1} \in B) = \frac{\sum_{j=0}^{n} \mathbb{1}(I_j \cap B \neq \emptyset)}{n+1} \tag{2.4.3}$$

where $\mathbb{1}(\cdot)$ is the indicator function. The lower probability is equal to the summation of the probabilities assigned to intervals I_j that are completely within the set B. For the upper probability, we sum up all probabilities assigned to intervals that intersect with the set B; so all I_j such that $I_j \cap B \neq \emptyset$.

While the assumption $A_{(n)}$ inherently offers a predictive probability for only a single future observation, its applicability can be expanded to m future observations, denoted as random variables $X_{n+1}, \ldots, X_{n+m}$. The data and future observations are linked by Hill's assumption $A_{(n)}$, or more precisely by applying consecutively $A_{(n)}, A_{(n+1)}, \ldots, A_{(n+m-1)}$ [21], which together can be called the $A_{(\cdot)}$ assumptions. Then the assumptions are a post-data version of a finite exchangeability assumption for $n+m$ random variables. Based on$A_{(\cdot)}$ assumptions, all possible orderings of the n data observations and the m future observations are equally likely, in which the n data observations and m future observations are not distinguished from each other [21]. In the NPI framework, exchangeability means that all possible orderings of real-valued data are equally likely before any values are observed. Let S_{ji} represent the number of the m future observations that fall within the interval $I_j = (x_{j-1}, x_j)$, so that $S_j = \#\{X_{n+i} \in I_j, i = 1, \ldots, m\}$, then assuming $A_{(\cdot)}$, we have [21]

$$P(\bigcap_{j=1}^{n+1} \{S_j = s_j\}) = \binom{n+m}{m}^{-1} \tag{2.4.4}$$

where s_j, for $j = 1, \ldots, n+1$, are any non-negative integers with $\sum_{j=1}^{n+1} s_j = m$. Equation (2.4.4) implies that all $\binom{n+m}{m}$ orderings of m future observations among the n observations are equally likely. For example, when $n = 3$ and $m = 5$, there are $\binom{3+5}{5} = 56$ possible different orderings of the m future observations among the n real data observations.

2.5 NPI for circular data

Circular data refers to information measured on a circular scale, signifying that the values follow a cyclical pattern [27]. In contrast to linear data, which is quantified on a real line with defined endpoints, circular data forms a loop with no clear starting or ending point. Common examples of circular data include compass directions and time. Circular data is prevalent across various research fields, including medical sciences, personality measurement, political science, sociology, and education [27].

In the case of circular data, $A_{(n)}$ in its standard form is not suitable, as the data is not represented on a real line. Due to this limitation, and linking back to the exchangeability of $n + 1$ observations, the assumption is called circular $A_{(n)}$ and is denoted by $\textcircled{A}_{(n)}$ [20]. Consider ordered circular data $x_1 < x_2 < \ldots < x_n$ with n intervals called $C_j = (x_j, x_{j+1})$ for $j = 1, \cdots, n-1$ and $C_n = (x_n, x_1)$. The assumption $\textcircled{A}_{(n)}$ for future random quantity X_{n+1} is [20]:

$$P(X_{n+1} \in C_j) = \frac{1}{n} \quad \text{for } j = 1, 2, \ldots, n. \tag{2.5.1}$$

The $\textcircled{A}_{(n)}$ is a post-data assumption related to the exchangeability assumption for circular data. The lower and upper probabilities for single future observations with circular data $\textcircled{A}_{(n)}$ are presented by Coolen [20]. The probabilities for X_{n+1} as defined by $\textcircled{A}_{(n)}$ directly correspond to lower and upper probabilities for events of the form $X_{n+1} \in B$, where B is a segment of the circle [7]. The lower probability $\underline{P}(X_{n+1} \in B)$ is obtained by adding the probability masses related to intervals C_j within B. The upper probability $\overline{P}(X_{n+1} \in B)$ is obtained by adding the probability masses related to all intervals C_j which have a non-empty intersection with B [20].

From assumption of $\textcircled{A}_{(n)}$ given as the predictive probability for single future values based on Coolen [20], this can be extended for m future observations. The extension to m future observations, denoted as random quantities $X_{n+1}, \ldots, X_{n+m}$, involves the simultaneous assumptions $\textcircled{A}_{(n)}, \textcircled{A}_{(n+1)}, \ldots, \textcircled{A}_{(n+m-1)}$, which together are denoted by $\textcircled{A}_{(\bullet)}$. The $\textcircled{A}_{(\bullet)}$ assumptions imply that all possible orderings of m future observations among n data observations are equally likely. Let S_j denote the number of observations into category C_j. Then the $\textcircled{A}_{(\bullet)}$ assumption lead to [20]

$$P(\bigcap_{j=1}^{n}\{S_j = s_j\}) = \binom{n+m-1}{m}^{-1} \tag{2.5.2}$$

where s_j, for $j = 1, \ldots, n$, are any non-negative integers with $\sum_{j=1}^{n} s_j = m$. Equation (2.5.2) implies that all $\binom{n+m-1}{m}$ possible orderings of m future observations among the n observations are an equally likely. Generally, in NPI, as discussed in Section 2.4, the NPI lower probability is derived by counting all the orderings for which the event of interest must hold, while the NPI upper probability is derived by counting all the orderings for which the event of interest can hold [7, 20]. In chapter 3, NPI is generalised for a 2×2 table data based on the $\textcircled{A}_{(n)}$ assumption.

2.6 NPI bootstrap

Coolen and Binhimd [11, 23] introduced nonparametric predictive inference bootstrap method, referred to as NPI-B. The NPI bootstrap method is used for predictions, unlike regular bootstrap methods, which are used for estimations. NPI-B involves creating $n + 1$ intervals between n ordered observations of the original data, and selecting one of these intervals at random. From the chosen interval, one observation is drawn and added to the original data, leading to a total of $n + 1$ observations. The new data are $n + 1$ observations, so $n + 2$ intervals are created and then the method is repeated to have the new $n + 2$ data. This process is repeated until one bootstrap sample of size m is obtained. The NPI-B algorithm for one-dimensional real-valued data on a finite (bounded) interval is as follows [11, 23]:

1. Take the data set of n real-valued observations.

2. Create $n + 1$ intervals based on the n original data set.

3. Choose randomly one of these intervals with equal probability.

4. Sample one future value from this selected interval.

5. Add that value to the data: increase n to $n + 1$.

6. Repeat Steps 2-4, now with $n + 1$ data, to get a further future value.

7. Do this m times to get a NPI bootstrap sample of size m, only the new sampled data are considered as the bootstrap sample.

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

The NPI-B algorithm can be adapted for circular data, allowing its use in the NPI method with 2×2 table data for inference. A detailed discussion on this is presented in Section 3.5. The NPI bootstrap method is used in this thesis for predictions with a focus on reproducibility, unlike regular bootstrap methods, which are used for estimations.

2.7 NPI sampling of orderings method

A sampling methodology based on sampling future orderings is proposed by Coolen and Marques [24] to overcome the computational limitations associated with large sample sizes. This technique satisfies the conditions of simple random sampling (SRS), ensuring each selection has an equal probability of being chosen and each selection is independent of the others. For simplicity, orderings are sampled with replacement, as a high number of orderings effectively eliminates discrepancies between sampling with and without replacement.

In this study, we utilize the sampling of orderings approach for circular data. A latent variable is used to represent observations on circular data. Denote the $\binom{n+m-1}{m}$ different orderings of m future observations among the n data observations on the circle by O_j for $j = 1, \ldots, \binom{n+m-1}{m}$. The ordering of O_j can be represented by $(s_1, \ldots, s_n)$, where s_i is the number of future observations that fall into the intervals $C_1, \ldots, C_n$, according to the ordering of O_j with $\sum_{i=1}^{n} s_i = m$. The orderings of future observations refer to the different possible ways future data points can be distributed among the existing n data points on the circle. For example, with $n = 3$ existing observations and $m = 2$ future observations, the number of possible orderings is $\binom{3+2-1}{2} = \binom{4}{2} = 6$, and one possible ordering could be $(s_1 = 1, s_2 = 0, s_3 = 1)$.

For the orderings, the sampling technique meets the SRS requirements. Each selection must have the same probability of being chosen, and each selection must be independent. By sampling with replacement, we simplify the process and maintain accuracy. Sampling orderings is done by selecting a vector of integers $\{r_2, \ldots, r_n\}$, where r_i represents the rank of the jth ordered observation in the combined data and future observations. Let $s_i^j = r_{i+1} - r_i - 1$ for $i = 1, \ldots, n$, with $r_1 = 1$ and $r_{n+1} = n + m + 1$. This process produces the jth sampled future ordering, ensuring that every possible ordering has an equal chance of being selected and is independent of other selections. The sampling of orderings approach for circular data will be used in chapters 3 and 4.

Confidence intervals for a single population proportion p are fundamental concepts in statistics. The Normal approximation assumes that the distribution of the sample proportion $\hat{p}$ is approximately Normal when the sample size n is large. When the sample size is large enough and $\hat{p}$is not close to 0, the Normal approximation method can be used [15]:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \tag{2.7.3}$$

where $z_{\frac{\alpha}{2}}$ is the $1 - \frac{\alpha}{2}$ quantile of the standard Normal distribution. In Chapter 4, we used confidence intervals based on the standard normal distribution for the approximate lower and upper probabilities for test reproducibility.

For small sample sizes and when p is close to 0, the Clopper–Pearson interval, also known as the 'exact' method, is preferred. The Clopper-Pearson interval, often referred to as the 'exact' method, does not rely on large-sample approximations and instead uses the exact binomial distribution to determine confidence bounds [63]. It is based on the cumulative probabilities of the binomial distribution, providing an exact distribution for a single population proportion p rather than an approximation. The Clopper-Pearson interval is computed under the assumption that the data follow a binomial distribution with the null hypothesis $H_0 : p = p_0$. This assumes that each trial is independent and that the probability of success p remains constant. The Clopper-Pearson confidence interval for p is derived by inverting two single-tailed binomial tests $H_0 : p = p_0$. The endpoints of the interval are solutions in p_0 to the equations [3, 63]:

$$\sum_{k=x}^{n} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \frac{\alpha}{2}$$

$$\sum_{k=0}^{x} \binom{n}{k} p_0^k (1-p_0)^{n-k} = \frac{\alpha}{2}$$

where x is the number of observed successes in a sample of size n. Note that the lower bound is 0 when $x = 0$ and the upper bound is 1 when $x = n$ [3, 63]. The confidence interval is [3, 63]:

$$\left[1 + \frac{n-x+1}{xF\left(2x, 2(n-x+1), 1-\frac{\alpha}{2}\right)}\right]^{-1} < p < \left[1 + \frac{n-x}{(x+1)F\left(2(x+1), 2(n-x), \frac{\alpha}{2}\right)}\right]^{-1} \tag{2.7.4}$$

Here, $F(a, b, c)$ represents the $1-c$ quantile of the F-distribution with a and b degrees of freedom. The Clopper-Pearson interval is computed under the assumption that the data follow a binomial distribution, aiming to provide exact coverage probabilities regardless of the true proportion p. The formula calculates the exact confidence interval for p by inverting the binomial cumulative distribution function, ensuring accurate coverage even with small sample sizes or extreme proportions. Alternatively, the lower endpoint can be described as the $\frac{\alpha}{2}$ quantile of a Beta distribution with parameters x and $n-x+1$, and the upper endpoint as the $1-\frac{\alpha}{2}$ quantile of a Beta distribution with parameters $x+1$ and $n-x$ [3]. In Chapter 3, we used the Clopper-Pearson interval for the approximate lower and upper probabilities.

2.8 NPI for reproducibility probability of statistical tests

The reproducibility of test outcomes is an important characteristic of practical statistics. The explicitly predictive nature of NPI provides a natural formulation for inference on reproducibility of tests. The reproducibility probability (RP) plays an important role in determining the reliability of statistical tests. Recently there has been substantial interest within the reproducibility probability (RP), where not only its estimation but also it is the particular definition and interpretation do not seem to be uniquely determined within the classical frequentist statistics framework as discussed in Section 2.3.

Coolen and Bin Himd [11, 23] introduced NPI for RP, denoted by NPI-RP, by considering some basic nonparametric tests such as the sign test, Wilcoxon's signed-rank test and the two-sample rank-sum test. For these inferences, NPI for Bernoulli quantities and for real-valued observations were used [7]. The NPI lower and upper reproducibility probabilities of the test are denoted by $\underline{RP}$ and $\overline{RP}$, respectively. As a measure of reproducibility, the test result is calculated for a predicted future sample that has the same size as the original sample ($m = n$). In the NPI-RP approach, the fundamental concept considers the $\binom{2n}{n}$ different orderings of n future real-valued observations among n data observations, where these orderings all have the same probability to occur. Let O_j for $j = 1, \ldots, \binom{2n}{n}$ represent the different orderings of the n future observations among the n data observations. Each ordering O_j can be represented by $(s_1^j, \cdots, s_{n+1}^j)$ for $i = 1, 2, ..., n+1$, where s_i^j is the number of future observations in $I_j = (x_{(i-1)}, x_{(i)})$ for $i = 1, \ldots, n+1$ with respect to ordering O_j such that $\sum_{i=1}^{n} s_i^j = n$ [11, 23]. In NPI-RP, all possible orderings of n future observations among n data observations are considered, given the observed data from the original test. On the future data sets the same test is conducted as it was on the original data, and the proportion of these that lead to the same conclusion is determined. Coolen and Alqifari [21] presented NPI-RP for two basic nonparametric tests based on order statistics. Marques et al. [56] investigated the reproducibility of the likelihood ratio test using NPI.

Coolen and Binhimd [11] used the NPI-bootstrap method to study reproducibility. They demonstrated how the NPI-bootstrap method avoids complicated calculations encountered with the NPI-RP method. A limitation of the NPI-RP method is that large sample sizes make computations impractical. Simkus et al. [67] employed the NPI-bootstrap method to assess the reproducibility of the t-test. Simulations were conducted to determine the reproducibility under both null and alternative hypotheses.

For larger samples, it becomes computationally challenging to consider all orderings to calculate NPI-RP. To avoid this difficulty, two NPI methods are used: NPI bootstrap and NPI sampling of orderings. The NPI bootstrap method gives a point estimate of reproducibility probabilities, while the sampling of orderings method estimates the lower and upper reproducibility probabilities. A sampling methodology based on sampling orders is proposed by Coolen and Marques [24] for overcoming the computational limitations associated with large sample sizes.

Chapter 3

NPI Circular Data Method for 2×2 Contingency Tables

3.1 Introduction

Nonparametric predictive inference (NPI) has been used in a variety of statistical applications, providing imprecise probabilities for both single and multiple future observations. In this chapter, NPI is generalized for inferences from 2×2 table data based on the assumption $\textcircled{A}_{(n)}$. The motivation behind this work is to extend the applicability of NPI to 2×2 table data, which is a common format in many fields. The reason for using the assumption $\textcircled{A}_{(n)}$ is that the categories are adjacent to each other. For more explanation, see Section 3.2.

This chapter is organized as follows: Section 3.2 introduces NPI for circular data as a method for inferences based on 2×2 table data. Sections 3.3 and 3.4 present the NPI lower and upper probabilities for events involving multiple future observations. Section 3.5 presents NPI-bootstrap for circular data as a computational method for 2×2 table data. The chapter concludes with final remarks in Section 3.6.

3.2 NPI for circular data as method for 2×2 tables

In this section, we introduce a new method for inference from 2×2 tables based on NPI for circular data. The foundation of method is based on the $\textcircled{A}_{(n)}$ assumption,

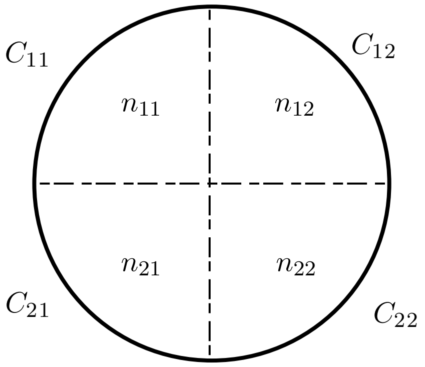


Figure 3.1: Circular data representation for 2×2 table data

a variant of Hill's assumption $A_{(n)}$, discussed in Section 2.5. For simplicity, we use the $\textcircled{A}_{(n)}$ assumption for 2×2 tables to make each cell ij, referred to by C_{ij}, corresponding to category ij for $i, j \in 1, 2$. The $\textcircled{A}_{(n)}$ assumption is used because the categories are adjacent to each other. For 2×2 tables , we employ a latent variable in a circular representation to represent the categories. This assumption helps to compute the lower and upper probabilities for X_{n+1}; this will be discussed later in this section.

It should be noted that each category C_{ij} is divided into n_{ij} equal-sized segments. Each observation is represented by a single segment of the probability wheel, where a segment is the area between two lines from the center to its circumference. According to the circular $\textcircled{A}_{(n)}$ assumption given in Equation (2.5.1), the probability of the next future observation falling into any given segment is $\frac{1}{n}$. Inferences about the next observation are based on a latent variable representation using a probability wheel.

For 2×2 table data, we can calculate NPI lower and upper probabilities without adding any further additional assumptions. Let n_{ij} represent the numbers of observations in C_{ij}, such that $\sum_{i,j} n_{ij} = n$ for $i, j \in \{1, 2\}$ (see Figure 3.1). Based on the $\textcircled{A}_{(n)}$ assumption, the NPI lower and upper probabilities for the event $X_{n+1} \in C_{ij}$, referred to by $\underline{P}_{ij}$ and $\overline{P}_{ij}$, are computed by Equations (3.2.1) and (3.2.2), respectively. Hence, the lower probability $\underline{P}(X_{n+1} \in C_{ij})$ is derived by summing only the probability masses assigned to segments entirely within C_{ij}, excluding any neighboring segments. Conversely, the upper probability $\overline{P}(X_{n+1} \in C_{ij})$ is obtained by summing all probability masses assigned to segments within C_{ij}, along with the probability mass for the neighboring segment of the category C_{ij}. The lower and upper probabilities for the event $X_{n+1} \in C_{ij}$ for $i, j \in \{1,2\}$ can be derived using NPI for circular data as :

$$\underline{P}_{ij} = \underline{P}(X_{n+1} \in C_{ij}) = \frac{n_{ij} - 1}{n} \quad \text{for} \quad i, j \in \{1,2\} \tag{3.2.1}$$

where $n_{ij} > 0$ for $i, j \in \{1,2\}$

$$\overline{P}_{ij} = \overline{P}(X_{n+1} \in C_{ij}) = \frac{n_{ij} + 1}{n} \quad \text{for} \quad i, j \in \{1,2\} \tag{3.2.2}$$

NPI Lower Probabilities ($\underline{P}_{ij}$) calculated considering only segments completely within the category C_{ij}, excluding any adjacent segments. NPI Upper Probabilities ($\overline{P}_{ij}$) calculated by including all segments within category C_{ij}, plus one adjacent segment. This accounts for the possibility that the next observation could fall into an adjacent category due to the circular nature of the data.

Example 3.2.1. Consider the data in Table 3.1 with $n = 7$ observations, and $m = 1$. Visualizing these data on a circular probability wheel, the wheel is divided into 7 equal segments, each representing an observation with a probability of $\frac{1}{7}$. Applying the Ⓐ$_{(n)}$ assumption, we calculate the NPI lower and upper probabilities for the event that the next observation for each category as follows. For C_{11} with 2 observations, the lower probability is $\underline{P}_{11} = \frac{2-1}{7} = \frac{1}{7}$, accounting for segments entirely within C_{11}, while the upper probability is $\overline{P}_{11} = \frac{2+1}{7} = \frac{3}{7}$, which includes one adjacent segment. For C_{12} with 1 observation, the lower probability is $\underline{P}_{12} = \frac{1-1}{7} = 0$, indicating no segments entirely within C_{12}, and the upper probability is $\overline{P}_{12} = \frac{1+1}{7} = \frac{2}{7}$, including one adjacent segment. Similarly, for C_{21} and C_{22}, each with 2 observations, the lower probabilities are $\underline{P}_{21} = \underline{P}_{22} = \frac{2-1}{7} = \frac{1}{7}$, and the upper probabilities are $\overline{P}_{21} = \overline{P}_{22} = \frac{2+1}{7} = \frac{3}{7}$. This example demonstrates how the lower probability of each category accounts only for the segments entirely within it, while the upper probability also includes one adjacent segment, reflecting the circular nature of the data and the Ⓐ$_{(n)}$ assumption.

Example 3.2.2. Responses for two independent samples of treatments at low and high doses are shown in Table 3.2 [48]. This data helps evaluate the effectiveness

	Treatment 1	Treatment 2	Total
Group 1	2	1	3
Group 2	2	2	4
Total	4	3	7

Table 3.1: Comparison of the effectiveness of two treatments

		Response		Total
		Success	Failure	
Variables	High	41	9	50
	Low	37	13	50
Total		78	22	100

Table 3.2: Response frequencies for low and high dose treatments.

of the treatment at different doses. The aim of this example is to show how to compute the NPI lower and upper probabilities for X_{101}. The NPI lower and upper probabilities in Table 3.2, with $n = 100$ and $m = 1$, can be computed based on Equations (3.2.1) and (3.2.2). The lower $\underline{P}_{11}$ and upper $\overline{P}_{11}$ probabilities for the event of interest ($X_{101} \in C_{11}$) are calculated as follows:

$$\underline{P}_{11} = P(X_{101} \in C_{11}) = \frac{41-1}{100} = 0.40$$

$$\overline{P}_{11} = P(X_{101} \in C_{11}) = \frac{41+1}{100} = 0.42$$

Thus, the lower and upper probabilities for the event ($X_{101} \in C_{11}$) are 0.40 and 0.42, respectively. This indicates the range within which we predict the probability of success for the high dose treatment in the next trial. Table 3.3 shows the results for the lower and upper probabilities for different cells.

$\underline{P}_{ij}$	$\underline{P}_{ij}$	$\overline{P}_{ij}$	$\overline{P}_{ij}$
0.40	0.08	0.42	0.10
0.36	0.12	0.38	0.14

Table 3.3: NPI lower and upper probabilities for the 2 × 2 contingency table.

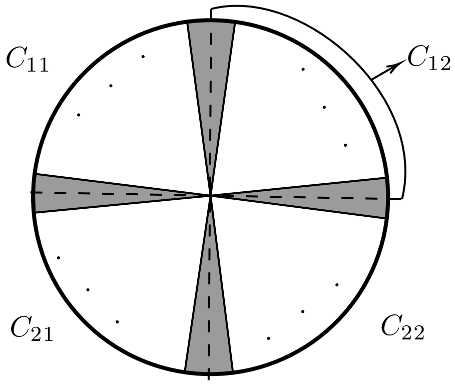


Figure 3.2: Illustration of segments between the two categories.

3.3 NPI lower probability for multiple future observations

In Section 2.5, inferences were presented for events involving multiple future observations based on $\textcircled{A}_{(\bullet)}$. In this section, we present the NPI lower probabilities for events involving multiple future observations based on $\textcircled{A}_{(\bullet)}$ for 2×2 tables. For 2×2 table data, the NPI method employs an assumed underlying latent variable model in which the categories are represented as intervals on a circular . This representation preserves the known ordering of the categories and facilitates the application of assumption $\textcircled{A}_{(\bullet)}$. The NPI lower probability for the future observation $X_{ij} = x_{ij}$ for $i, j \in \{1, 2\}$ is derived by counting all orderings where this event must hold. For $m \geq 2$ future observations, we use the notation X_{ij} to denote the number of these m future observations that fall into category C_{ij} for $i, j \in \{1, 2\}$, such that $X_{ij} = x_{ij}$. Therefore, $\sum_{i,j} X_{ij} = m$, representing the total number of future observations.

The total number of different arrangements of the original observations is equal to $\binom{n+m-1}{m}$, representing the different orderings of m future observations among

	Control	Treatment	Total
Male	3	2	5
Female	4	3	7
Total	7	5	12

Table 3.4: Distribution of participants in control and treatment groups based on gender.

the n data observations [20]. For the NPI lower probability, we only consider the segments that are completely contained within the category C_{ij}. Each category includes only $n_{ij} - 2$ segments because two segments are associated with adjacent categories, as shown in Figure 3.2, where the relevant segments are shaded. The number of different arrangements of x_{ij} future observations within this segment is equal to

$$\binom{x_{ij} + n_{ij} - 2}{x_{ij}} \tag{3.3.3}$$

The NPI lower probability for the event of interest, based on m future observations, is given by:

$$\underline{P}_{ij}\left(\bigcap_{i,j}\{X_{ij} = x_{ij}\}\right) = \frac{1}{\binom{n+m-1}{m}} \prod_{ij} \binom{x_{ij} + n_{ij} - 2}{x_{ij}} \tag{3.3.4}$$

where x_{ij}, for $j = 1, 2$, are non-negative integers with $\sum_{ij} x_{ij} = m$. Equation 3.3.4 defines the NPI lower probability for observing a specific distribution of m future observations across all categories in a 2×2 table. It is calculated by taking the product of the binomial coefficients for each category C_{ij}, which represent the possible arrangements of observations within those categories, and then dividing by the total number of possible arrangements of the m future observations among the existing n data .

Example 3.3.1. To illustrate, we consider the simulated data in Table 3.4 with $n = 12$ and $m = 2$ future observations such that $X_{11} = 1, X_{12} = 1, X_{21} = 0, X_{22} = 0$. For the this event, NPI lower probability by using Equation (3.3.4), is

$$\underline{P}(X_{11} = 1, X_{12} = 1, X_{21} = 0, X_{22} = 0) = \frac{\binom{3+1-2}{1}\binom{2+1-2}{1}\binom{4+0-2}{0}\binom{0+0-2}{0}}{\binom{12+2-1}{2}} = 0.026$$

Table 3.5 presents the lower probabilities for different events with different numbers of future observations. As the number of future observations increases, NPI lower probabilities decrease. From Table 3.5, it could be happened that all future observations $m = 5$ falls into one category, with a probability of 0.001373 for the event {5, 0, 0, 0}. Conversely, the event {0, 0, 2, 3}, where observations are spread across categories, has a higher probability of 0.005494, making it more likely. The probability is low because as the number of future observations increases, there are more ways they can be distributed, making specific distributions less likely. For example, having all future observations in one category is unlikely because they will likely be spread across multiple categories. Also, the initial data distribution affects these probabilities; categories that have very few observations in the current data are less likely to receive many future observations. Table 3.6 presents the NPI lower probabilities from Example 3.2.2 for $n = 100$ with different numbers of future observations. Similar to the results in Table 3.5, as the number of future observations increases, the NPI lower probabilities decrease.

3.4 NPI upper probability for multiple future observations

The NPI upper probability for the event $X_{ij} = x_{ij}$ is determined by considering all orderings where this event can hold. The upper probabilities for multiple future observations are discussed in Section 2.5 . This section discusses the general case of m future observations, where X_{ij} represents the number of future observations in C_{ij}, and $\sum_{ij} X_{ij} = m$ for $i, j \in \{1, 2\}$. The event of interest is $X_{11} = x_{11}$, $X_{12} = x_{12}$, $X_{21} = x_{21}$, and $X_{22} = x_{22}$. Based on $\textcircled{A}_{(\bullet)}$, the number of different arrangements of future m observations among n data observations is equal to $\binom{n+m-1}{m}$. Deriving an exact equation for the NPI upper probability is difficult because future observations can be placed in different segments between categories. Each way of arranging

$m=2$		$m=5$	
$\{x_{11}, x_{12}, x_{21}, x_{22}\}$	$\underline{P}$	$\{x_{11}, x_{12}, x_{21}, x_{22}\}$	$\underline{P}$
{2,0,0,0}	0.0384	{5,0,0,0}	0.001373
{0,2,0,0}	0.0128	{0,5,0,0}	0.000228
{0,0,2,0}	0.0769	{0,0,5,0}	0.004807
{0,0,0,2}	0.0384	{0,0,0,5}	0.001373
{1,1,0,0}	0.0256	{2,2,1,0}	0.002060
{1,0,1,0}	0.0769	{1,1,0,3}	0.001831
{1,0,0,1}	0.0512	{0,0,2,3}	0.005494
{0,1,1,0}	0.0384	{1,1,3,0}	0.004578
{0,1,0,1}	0.0257	{2,1,1,1}	0.004120
{0,0,1,1}	0.0769	{3,0,0,2}	0.002747

Table 3.5: NPI lower probabilities for multiple future observations, Example 3.3.1

these observations affects which category they belong to. We need to consider all possible orderings where the event of interest can occur. This requirement makes it difficult to formulate an exact equation for the NPI upper probability. Instead, an approximation can be derived using sampling of orderings. The approximation procedure involves two main steps: first, we use the sampling of orderings approach for circular data as discussed in Section 2.7. Then, for a given ordering of m future observations among the n data observations, we aim to verify whether the event of interest can occur or not.

Given n data observations with m future observations, the $\binom{n+m-1}{m}$ different orderings of all these observations are equally likely on the circle, where each ordering O_j for $j = 1, \dots, \binom{n+m-1}{m}$ can be represented by $(s_1, \dots, s_n)$, and s_i is the number of future observations that fall into C_{ij} for $i, j \in \{1, 2\}$ such that $\sum_{i=1}^{n} s_i = m$. Consider the data with n_{ij} observations in C_{ij} categories for $i, j \in \{1, 2\}$, such that n_{ij} in C_{ij} for $i, j \in \{1, 2\}$, so $\sum_{i,j} n_{ij} = n$ for $i, j \in \{1, 2\}$. Let X_{ij} be the number out of these m future observations that fall into C_{ij} for $i, j \in \{1, 2\}$, such that $X_{ij} = x_{ij}$ for $i, j \in \{1, 2\}$.

$m = 5$		$m = 10$	
$\{x_{11}, x_{12}, x_{21}, x_{22}\}$	$\underline{P}$	$\{x_{11}, x_{12}, x_{21}, x_{22}\}$	$\underline{P}$
{5,0,0,0}	0.011800	{10,0,0,0}	0.000192
{0,5,0,0}	0.000008	{2,1,6,1}	0.008302
{0,0,5,0}	0.007155	{1,1,1,7}	0.000008
{0,0,0,5}	0.000047	{1,1,5,3}	0.001797
{2,1,2,0}	0.047508	{2,2,4,2}	0.004442
{1,1,3,0}	0.029354	{2,5,2,1}	0.000121
{3,0,1,1}	0.053928	{3,2,2,3}	0.002349
{0,2,3,0}	0.003302	{5,2,2,1}	0.007328
{2,1,1,1}	0.030816	{6,1,2,1}	0.012214
{2,1,0,2}	0.005564	{4,1,1,4}	0.001137

Table 3.6: NPI lower probabilities for multiple future observations, Example 3.2.2

Given a vector $(s_1, \ldots, s_n)$ with $\sum_{j=1}^{n} s_j = m$, the m future observations can fall entirely within specific categories or across neighboring segments. We define s^{ij} as the number of segments completely within category C_{ij} for $i, j \in \{1, 2\}$, as shown in Figure 3.3. Specifically, for category C_{11}, the segments are given by $s^{11} = \sum_{i=1}^{n_a - 1} s_i$, while for C_{12}, they are given by $s^{12} = \sum_{i=n_a+1}^{n_b - 1} s_i$. Similarly, for C_{22}, the segments are given by $s^{22} = \sum_{i=n_b+1}^{n_c - 1} s_i$, and for C_{21}, the segments are given by $s^{21} = \sum_{i=n_c+1}^{n_d - 1} s_i$. Here, the indices a, b, c, and d are defined as $n_a = n_{11}$, $n_b = n_{1+}$, $n_c = n_{11} + n_{12} + n_{22}$, and $n_d = n$. For example, with $n_{11} = 3$, $n_{12} = 2$, $n_{22} = 2$, and $n_{21} = 3$, the segments are $s_{11} = s_1 + s_2$, $s_{12} = s_4$, $s_{22} = s_6$, and $s_{21} = s_8 + s_9$. This means that category C_{11} includes the first two segments, C_{12} includes the fourth segment, C_{22} includes the sixth segment, and C_{21} includes the eighth and ninth segments. Additionally, the segments s_3, s_5, s_7, and s_{10} are neighboring segments in the circle.

We can represent the segments of the circle between two neighboring observations in neighboring categories such that s_a represents the segment between the categories C_{11} and C_{12}, s_b corresponds to the segment between categories C_{12} and C_{22}, s_c denotes the segment between categories C_{22} and C_{21}, and finally, s_d represents the

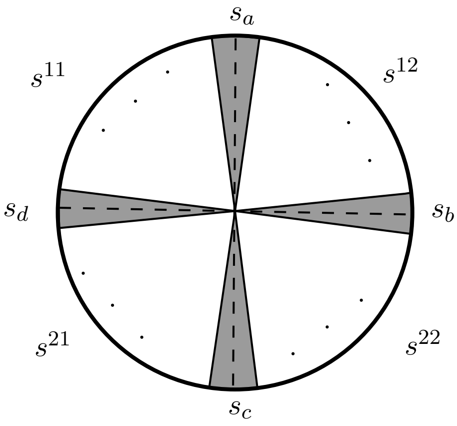


Figure 3.3: Illustration of s^{ij}, where s^{ij} is the number of segments that are completely in category C_{ij}, and the shaded areas s_a, s_b, s_c and s_d are the segments between any two categories.

segment between categories C_{21} and C_{11}. The values s_a, s_b, s_c, and s_d can be any non-negative integers. Let s_* be the neighboring segment between any two categories and s_*^{ij} be the number of future observations that are located in the segment $*$ from category C_{ij}. We can represent the segments as follows:

$$
\begin{aligned}
s_a &= s_a^{11} + s_a^{12}, \text{ for } s_a^{11}, s_a^{12} \in \{0, \dots, s_a\} \\
s_b &= s_b^{12} + s_b^{22}, \text{ for } s_b^{12}, s_b^{22}, \in \{0, \dots, s_b\} \\
s_c &= s_c^{22} + s_a^{21}, \text{ for } s_c^{22}, s_c^{21} \in \{0, \dots, s_c\} \\
s_d &= s_d^{11} + s_d^{21}, \text{ for } s_d^{11}, s_d^{21} \in \{0, \dots, s_d\}
\end{aligned}
$$

In the segments of categories C_{ij} for $i, j \in \{1, 2\}$, we investigate how the future observations are assigned based on the sampled of ordering for $i, j \in \{1, 2\}$. For category C_{11}, there are x_{11} future observations distributing between the segments, which may be inside the category or in the neighboring segments. This means that $x_{11} = s^{11} + s_a^{11} + s_d^{11}$. In the categories C_{12}, C_{22}, and C_{21}, it is clear that $x_{12} = s^{12} + s_a^{12} + s_b^{12}$, $x_{22} = s^{22} + s_b^{22} + s_c^{22}$, and $x_{21} = s^{21} + s_d^{21} + s_c^{21}$, respectively. It is possible to have some future observations in the neighboring segments, but we have no information for which categories they belong to. To overcome this complexity, Algorithm 1 will be used. Algorithm 1 checks whether the event $X_{ij} = x_{ij}$ can occur based on sampled of orderings. The inputs include s^{ij}, x_{ij}, s_a, s_b, s_c, and s_d. The outputs of the Algorithm 1 indicates whether the event $X_{ij} = x_{ij}$ can occur or not.

The Algorithm 1 checks whether the event $X_{ij} = x_{ij}$ occurs for a given ordering of future observations among the data observations. At first, it checks if $s^{ij} > x_{ij}$ for any $i, j \in \{1, 2\}$; if so, the event is deemed negligible and cannot hold. If $s^{ij} = x_{ij}$, the event must hold, and all segments between categories are set to zero. When $X_{ij} = x_{ij}$ can potentially hold, the algorithm investigates each category individually: for C_{11}, if $s^{11} < x_{11}$, it distributes the remaining $x_{11} - s^{11}$ observations between s^{11}_d and s^{11}_a; similarly, for C_{12}, C_{22}, and C_{21}, it allocates the excess observations to their respective neighboring segments s^{12}_a and s^{12}_b for C_{12}, s^{22}_b and s^{22}_c for C_{22}, and s^{21}_d and s^{21}_c for C_{21}. After evaluating these conditions, the algorithm categorizes each ordering: if Step1 is satisfied, the event cannot occur; if Step2 is satisfied, the event must occur and increments the count for the NPI lower probability; if Step3 is satisfied, it increments the count for the NPI upper probability. By performing this procedure n^* times, the algorithm approximates the NPI lower probability by Count the total number of times the event must hold and divide by n^* , and approximates the NPI upper probability by count the total number of times the event can hold and divide by n^*.

To measure the effectiveness of the algorithm in providing an approximation for $\overline{P}_{ij}$, this algorithm will be used to approximate the NPI lower probability and compared to the exact value according to Equation (3.3.4). In Example 3.4.1, approximation are provided for both NPI lower and upper probabilities based on the Algorithm 1, along with the exact results for the NPI lower probability, which are based on Equation (3.3.4). The comparison gives insights on how Algorithm 1 is good to present approximations for the NPI lower probabilities; therefore, Algorithm 1 can be used to approximate the NPI upper probability. This defines the NPI lower probability $\underline{P}\{x_{11}, x_{12}, x_{21}, x_{22}\}$ and upper probability $\overline{P}\{x_{11}, x_{12}, x_{21}, x_{22}\}$, corresponding to the event where $X_{11} = x_{11}, X_{12} = x_{12}, X_{21} = x_{21}$, and $X_{22} = x_{22}$.

Example 3.4.1. Consider the data in Table 3.7 with $n = 40$ observations, and $m = 15$ future observations such that $X_{11} = 3$, $X_{12} = 4$, $X_{21} = 3$, $X_{22} = 5$. In this example, we approximate NPI lower and upper probabilities based on the Algorithm 1.

Algorithm 1 Checking the occurrence of the event $X_{ij} = x_{ij}$

and approximate NPI lower and upper probabilities .

```
1: If $s^{ij} > x_{ij}$ for $i, j \in \{1, 2\}$, then this situation will be neglectable.
2: If $s^{ij} = x_{ij}$ for $i, j \in \{1, 2\}$, then the segment between any two categories will
   be equal to 0.
3: If $X_{ij} = x_{ij}$ can be true and hold, the following situations should be investigated.

     (i) If $s^{11} < x_{11}$, then $x_{11} - s^{11} = s_d^{11} + s_a^{11}$, where $s_d^{11} \in \{0, \ldots, s_d\}$ and
         $s_a^{11} \in \{0, \ldots, s_a\}$.
    (ii) If $s^{12} < x_{12}$, then $x_{12} - s^{12} = s_a^{12} + s_b^{12}$, where $s_a^{12} = s_a - s_a^{11}$ for $s_a^{12} \in$
         $\{0, \ldots, s_a\}$ and $s_b^{12} = x_{12} - s^{12} - s_a^{12}$ for $s_b^{12} \in \{0, \ldots, s_b\}$.
   (iii) If $s^{22} < x_{22}$, then $x_{22} - s^{22} = s_b^{22} + s_c^{22}$, where $s_b^{22} = s_b - s_b^{12}$ for $s_b^{22} \in$
         $\{0, \ldots, s_b\}$ and $s_c^{22} = x_{22} - s^{22} - s_b^{22}$ for $s_c^{22} \in \{0, \ldots, s_c\}$.
    (iv) If $s^{21} < x_{21}$, then $x_{21} - s^{21} = s_d^{21} + s_c^{21}$, where $s_d^{21} = s_d - s_d^{11}$ for $s_d^{21} \in$
         $\{0, \ldots, s_d\}$ and $s_c^{21} = s_c - s_c^{22}$ for $s_c^{21} \in \{0, \ldots, s_c\}$.

     1. If Step 1 is satisfied, then the event of interest cannot be true.
     2. If Step 2 is satisfied, then the event of interest must be true. Count the
        total number of times the event must hold and divide by $n^*$; this gives the
        approximate NPI lower probability $\underline{P}$.
     3. If Step 3 is satisfied, then the event of interest can be true. Count the
        total number of times the event can hold and divide by $n^*$; this gives the
        approximate NPI upper probability $\overline{P}$.
```

	Treatment 1	Treatment 2	Total
Group 1	11	9	20
Group 2	10	10	20
Total	21	19	40

Table 3.7: Comparison of the effectiveness of two treatments

Table 3.8 presents the approximations of the NPI lower $\underline{P}\{3,4,3,5\}$ and upper $\overline{P}\{3,4,3,5\}$ probabilities for the event $X_{11} = 3$, $X_{12} = 4$,$X_{21} = 3$, $X_{22} = 5$ with sampling of orderings of size n^*. When $n^* = 100$, NPI lower probability for the event $X_{11} = 3$, $X_{12} = 4$,$X_{21} = 3$, $X_{22} = 5$ is equal to 0, then as n^* increases to 500 and 1000, it becomes 0.0020 and 0.0010, respectively. From $n^* = 2000$ to $n^* = 200,000$, the approximate NPI lower probabilities are nearly identical to the exact lower probability, which is equal to 0.0018 and calculated by Equation (3.3.4). When $n^* = 100$, the NPI upper probability is equal to 0.0200, then as n^* increases to 500 and 1000, it becomes 0.0280 and 0.0240, respectively. From $n^* = 2000$ to $n^* = 200,000$, the NPI upper probabilities are nearly identical. This note leads to choose $n^* = 2000$ because of having stability in the approximations. For the approximations of the NPI lower and upper probabilities, 95% confidence intervals are provided using the Clopper–Pearson interval in Equation (2.7.4), and it is clear and sensible that the confidence intervals become narrower as n^* increases. Here, our focus is only on computing the values for the lower and upper bounds, while simply noting that the confidence interval can be applied. When the NPI lower and upper probability approximations are small, the Clopper–Pearson interval provides accurate and reliable confidence bounds, effectively addressing the challenges of extreme probabilities. From the presentation of Table 3.8, the algorithm with using sampling of orderings presents nearly identical results to the exact NPI lower probability when $n^* \geq 2000$. This gives insights that the algorithm can give good results for the NPI upper probability as $n^* \geq 2000$

It is worth to see the repetition of using the algorithm with sampling of orderings to approximate the NPI lower and upper probabilities for the event $X_{11} = 3$, $X_{12} = 4$, $X_{21} = 3$, $X_{22} = 5$. Figures 3.4 and 3.5 presents histograms for 100 approximations of NPI lower and upper probabilities for the event $X_{11} = 3$, $X_{12} = 4$, $X_{21} = 3$, $X_{22} = 5$ with sampling orders $n^* = 10,000$ and $100,000$, respectively. Those figures show that the NPI lower probability approximations are mostly near to 0.0018. For the NPI upper probability, most approximations are between 0.024 and 0.028. The ranges of approximations become smaller as the sampling of orderings n^* increases to 100,000. This leads to more evidence that the NPI lower and upper probabilities are approximately equal to 0.0018 and 0.026, respectively.

Table 3.9 presents the approximations of NPI lower and upper probabilities based on Example 3.4.1 with event $X_{11} = 5$, $X_{12} = 3$, $X_{21} = 2$, and $X_{22} = 5$ using sampling orderings of size n^*. When $n^* = 100$, the NPI lower probability for the event$X_{11} = 5$,$X_{12=} = 3$, $X_{21} = 2$, and $X_{22} = 5$ is equal to 0, then as n^* increases to 500 and 1000, it becomes 0.0014 and 0.0012, respectively. From $n^* = 2000$ to $n^* = 200,000$, the approximate NPI lower probabilities are nearly identical with the exact lower probability, =0.0017, which is calculated by Equation (3.3.4). When $n^* = 100$, the NPI upper probability is equal to 0.0190, then as n^* increases to 500 and 1000, it becomes 0.0180 and 0.0320, respectively. From $n^* = 2000$ to $n^* = 200,000$, the approximate NPI upper probabilities are nearly identical. For the approximations of NPI lower and upper probabilities, 95% confidence intervals, and it is clear that the confidence intervals become narrower as n^* increases.

n^*	$\underline{P}\{3,4,3,5\}$	95% CI	$\overline{P}\{3,4,3,5\}$	95% CI
100	0	(0, 0.0362)	0.0200	(0.0024, 0.0703)
500	0.0020	(0.0001, 0.0110)	0.0280	(0.0153, 0.0465)
1000	0.0010	(0.0002, 0.0055)	0.0240	(0.0154, 0.0355)
2000	0.0016	(0.0003, 0.0045)	0.0265	(0.0199, 0.0345)
5000	0.0018	(0.0008, 0.0034)	0.0264	(0.0221, 0.0312)
10,000	0.0016	(0.0009, 0.0025)	0.0254	(0.0224, 0.0286)
20,000	0.0020	(0.0009, 0.0026)	0.0242	(0.0221, 0.0264)
100,000	0.0017	(0.0014, 0.0019)	0.0254	(0.0244, 0.0263)
200,000	0.0018	(0.0016, 0.0019)	0.0262	(0.0255, 0.0269)

Table 3.8: Approximations of NPI lower and upper probabilities for the event $X_{11} = 3$, $X_{12} = 4$, $X_{21} = 3$, $X_{22} = 5$, Example 3.4.1

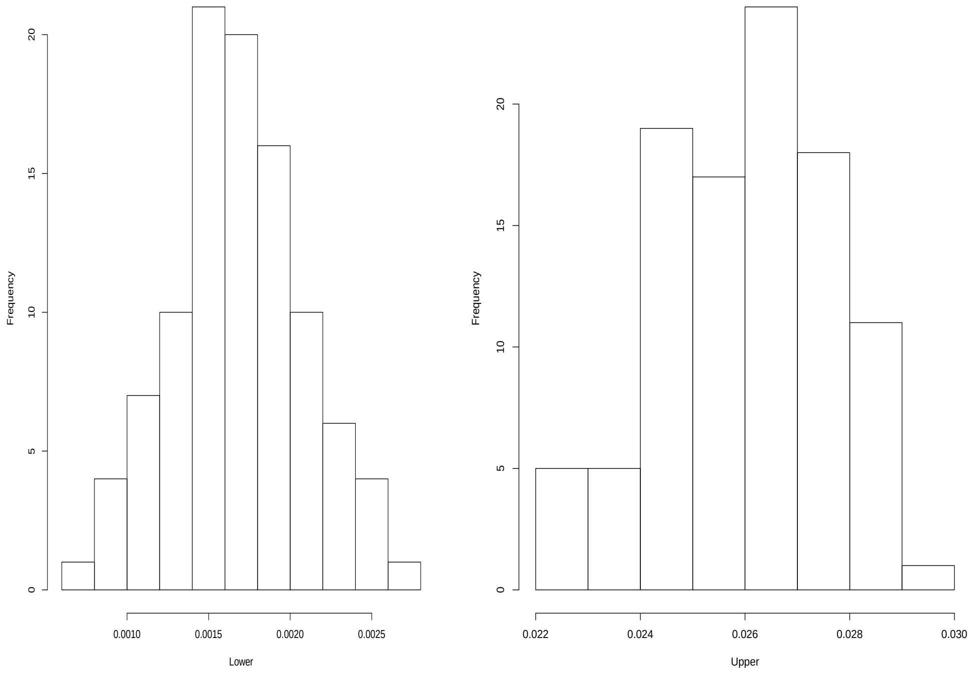


Figure 3.4: The approximations of NPI lower(left) and upper (right)probabilities for for the event$X_{11} = 3$, $X_{12} = 4$, $X_{21} = 3$, $X_{22} = 5$ with $n^* = 10,000$ and repetition for 100 times.

n^*	$\underline{P}$\{5,3,2,5\}	95% CI	$\overline{P}$\{5,3,2,5\}	95% CI
100	0	(0, 0.0385)	0.0190	(0.0021, 0.0688)
500	0.0014	(0.0003, 0.0100)	0.0180	(0.0082, 0.0338)
1000	0.0012	(0.0002, 0.0059)	0.0320	(0.0219, 0.0448)
2000	0.0015	(0.0006, 0.0043)	0.0240	(0.0177, 0.0316)
5000	0.0016	(0.0008, 0.0031)	0.0210	(0.0172, 0.0253)
10,000	0.0014	(0.0007, 0.0023)	0.0210	(0.0182, .02400)
20,000	0.0013	(0.0008, 0.0019)	0.0200	(0.0181, 0.0220)
100,000	0.0016	(0.0014, 0.0018)	0.0210	(0.0201, 0.0219)
200,000	0.0017	(0.0015, 0.0018)	0.0220	(0.0213, 0.0226)

Table 3.9: Approximations of NPI lower and upper probabilities for for the event$X_{11} = 5$, $X_{12} = 3$, $X_{21} = 2$, $X_{22} = 5$, Example 3.4.1

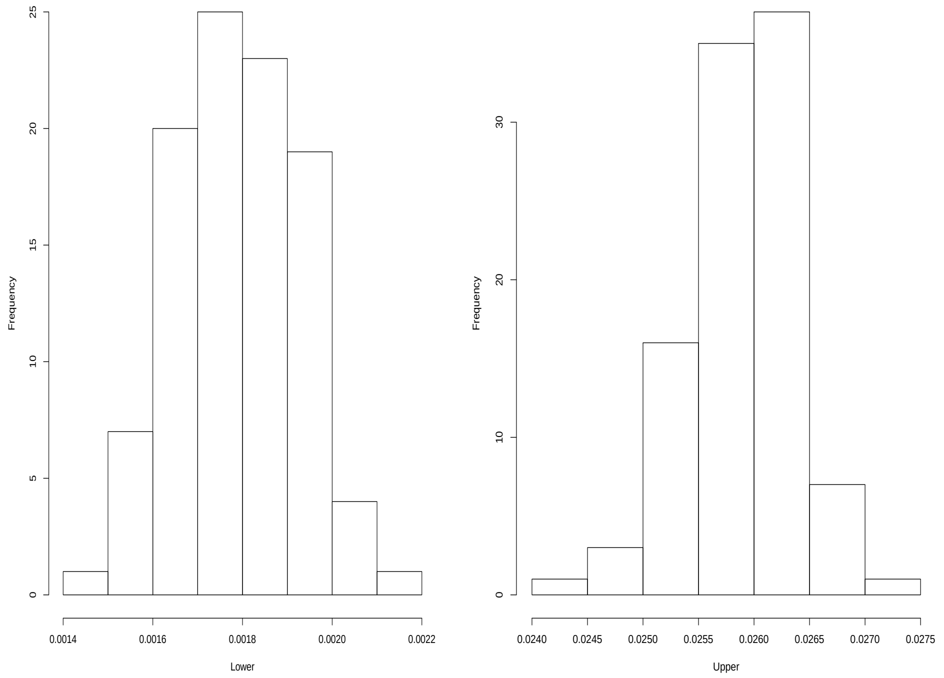


Figure 3.5: The approximations of NPI lower(left) and upper (right)probabilities for for the event$X_{11} = 3$, $X_{12} = 4$, $X_{21} = 3$, $X_{22} = 5$ with $n^* = 100,000$ and repetition for 100 times.

3.5 NPI-bootstrap for 2×2 tables

To enable computation in some NPI scenarios, Coolen and Binhimd [11, 23] introduced the NPI-bootstrap method, indicated by NPI-B, and discussed in Section 2.6. The NPI-B algorithm can be adapted for circular data, making it applicable for inferences in our NPI method for 2×2 table data. NPI-B is based on $\textcircled{A}_{(\bullet)}$, which was discussed in Section 2.5, and it is consistent with the idea that all orderings of future observations are equally likely. The algorithm of NPI bootstrap depends on the n segments created by the original data. For the NPI-B method, one segment is uniformly sampled, and the sampled segment is added to the original n segments. This leads to $n + 1$ segments. From the $n + 1$ segments, one segment is sampled uniformly and added, leading to $n + 2$ segments. This process is repeated m times to generate one NPI bootstrap sample of size m. The NPI-B for 2×2 tables as circular representation algorithm is as follows:

Algorithm 2 The NPI-B for 2×2 tables

1. The circle is divided into n segments based on the original observations. Each segment corresponds to a specific category and the segments between neighboring categories are considered.
2. One segment is sampled uniformly from the n segments and added to the original segments, leading to $n + 1$ segments.
3. Repeat step 2, now with $n + 1$ segments, to get $n + 2$ segments.
4. Repeat this process m times to get an NPI bootstrap sample of size m.
5. Count how many generated segments belong to each category and use these counts to form a new 2×2 table.
6. Repeat all steps B times to obtain B NPI bootstrap samples, each of size m, where B is any chosen integer value.

In this NPI-B algorithm, special attention be given to Step 2. When selecting one neighboring segment, which is between any two categories, the segment will be assigned to one category with a probability based on the number of observations in those categories in the original data. Each segment between categories is randomly assigned to one of the two categories based on a predefined probability. Segment s_a can be assigned to either C_{11} or C_{12} with probabilities $\frac{n_{11}}{n_{1+}}$ and $\frac{n_{12}}{n_{1+}}$, respectively. Segment s_b can be assigned to either C_{12} or C_{22} with probabilities $\frac{n_{12}}{n_{+2}}$ and $\frac{n_{22}}{n_{+2}}$, respectively. Segment s_c can be assigned to either C_{21} or C_{22} with probabilities $\frac{n_{21}}{n_{2+}}$ and $\frac{n_{22}}{n_{2+}}$, respectively. Finally, segment s_d can be assigned to either C_{21} or C_{11} with probabilities $\frac{n_{21}}{n_{+1}}$ and $\frac{n_{11}}{n_{+1}}$, respectively. These methods will be used in Chapters 4, 5, and 6.

3.6 Concluding remarks

In this chapter, NPI has been presented for circular data as a method for inference based on 2×2 table data. The NPI lower and upper probabilities are computed for one future observation based on $\textcircled{A}_{(n)}$ and for multiple future observations based on $\textcircled{A}_{(\bullet)}$. For one and multiple future observations, the NPI lower probabilities can be derived by exact formulas. In case of one future observation, it is simple to compute the NPI upper probability by an exact formula, but for multiple future observations, an exact formula for the NPI upper probability has not been derived, so an algorithm has been proposed to find an approximation.

For multiple future observations, the NPI lower probabilities are approximated by the algorithm and compared to the results computed by the exact formula. It is found that the approximations are nearly identical to the exact results as the sampling of orderings n^* is equal to 2000 or greater. This comparison is conducted to show how good the algorithm with sampling of ordering is to approximate the NPI lower probabilities, and due to the good results, the algorithm can be used to approximate the NPI upper probabilities. The NPI-B is a computational version of the NPI adapted for circular data, which used for inferences with 2×2 table data, and will be used in the following chapter.

Chapter 4

NPI Reproducibility of Tests based on 2×2 Contingency Tables

4.1 Introduction

The test reproducibility probability results is an important aspect of practical statistics. Test reproducibility probability is the probability that the hypothesis test outcome will be the same if an experiment is repeated in the same way as the original experiment. The definition of reproducibility and various methods were discussed in Section 2.3.

This chapter presents an investigation into the reproducibility of statistical hypothesis tests based on 2×2 tables. The tests covered are tests of independence, Fisher's exact test, and McNemar's test. Given its predictive nature, NPI is well-suited for inferences about reproducibility. The NPI approach provides a natural solution to the test reproducibility problem because it is fundamentally predictive. Test reproducibility is naturally considered a predictive inference problem because it involves estimating the probability that the test outcome will remain the same in future applications [11]. The NPI lower and upper RPs for hypothesis tests based on 2×2 tables are computed using sampling of orderings for circular data. Furthermore, the NPI-bootstrap (NPI-B) is employed to evaluate RP values for hypothesis tests based on 2×2 table.

The main aim is to use NPI sampling of orderings and NPI-B to assess the reproducibility of statistical hypothesis tests based on 2 × 2 tables. The explicitly predictive nature of NPI and NPI-B provides an appropriate formulation for inferring reproducibility. When the n data observations from the first test are used, the NPI-RP approach calculates a lower and upper probability that the test statistic will reject or not reject the null hypothesis, based on all possible order of n future observations and n data observations. Instead of considering all possible order, Coolen and Marques [24] proposed sampling future orders to approximate NPI-RP for larger sample sizes. We investigate the sampling of the orderings of future data among the observed data for circular data to approximate the lower and upper reproducibility probabilities for statistical hypothesis tests based on 2 × 2 tables.

The structure of this chapter is as follows: Section 4.2 introduces the NPI sampling of orderings and NPI bootstrap methods for reproducibility of the chi-square test and likelihood ratio “tests of independence. Section 4.3 explores reproducibility for McNemar's test using the same methods. Section 4.4 covers these approaches for the Fisher exact test. The chapter concludes with final remarks in the last section.

4.2 Reproducibility of tests of independence

This section introduces the use of NPI sampling of orderings and NPI-B for reproducibility of the chi-square test and likelihood ratio test of independence. It is common to use chi-square and likelihood ratio statistical tests to analyse 2×2 tables to assess to the two variables independence. In NPI-RP, reproducibility of tests is viewed from the perspective of prediction rather than estimation. This approach focuses explicitly on future observations and relies on few assumptions, which causes imprecision that can be quantified through lower and upper probabilities. Section 4.2.1 introduces approximations of NPI-RP for chi-square and likelihood ratio tests of independence, while Section 4.2.2 introduces the NPI-B approach to RP for these tests.

4.2.1 NPI-RP for tests of independence

We study the reproducibility of tests of independence using NPI sampling of orderings in this section. A fundamental concept of the NPI-RP approach is that it considers all possible orderings of n future observations among n existing data observations, each occurring with equal probability. The same test is applied to the future data sets as it was on the original data, and the proportion of these that reach the same conclusion as the original test is investigated. Deriving the NPI lower and upper reproducibility for a test is not analytically straightforward. Calculating NPI-RP for large datasets is computationally difficult because the number of possible orderings of future observations increases, leading to longer computation times. When the sample size is large, it is difficult to consider all the orderings required to calculate NPI-RP. Consequently, calculating NPI-RP for such large samples becomes complicated. As an alternative to considering all possible ordering, Coolen and Marques [24] propose sampling future of ordering instead of considering all possible ordering to approximate NPI-RP for larger sample sizes. In this work, we use sampling of orderings for circular data for chi-square and likelihood ratio tests. The aim of this section is to present approximating NPI-RP lower and upper reproducibility probability for the tests of independence. In this thesis, we approximate the NPI-RP lower and upper reproducibility probabilities through the sampling of orderings, rather than calculating the exact NPI-RP lower and upper reproducibility probabilities. For simplicity, this sampling method will be referred to as NPI-RP.

Considering $m = n$, the $\binom{2n-1}{n}$ represent different orderings of n future observations among the n data observations on the circle and denoting the number of orders, O_j for $j = 1, \ldots, \binom{2n-1}{n}$. Each ordering O_j can be represented by $(s_1, \ldots, s_n)$ for $i = 1, 2, ..., n$, where s_i is the number of future observations in C_{ij} for $i, j \in \{1, 2\}$ with respect to ordering O_j such that $\sum_{i=1}^{n} s_i = m$. Given vector $(s_1, \ldots, s_n)$ with $\sum_{i=1}^{n} s_i = m$, where the m future observations could be completely inside the categories or in the neighboring segments. Let s^{ij} is the number of segments that are completely in category C_{ij} for $i, j \in \{1, 2\}$. The segments between neighboring observations in different categories are denoted as s_a, s_b, s_c, and s_d(as explained in Section 3.4). Each segment can contain any number of observations from 0 to m.

Deriving a formula to generate the minimum and maximum values of the test test statistic for n future observations with a given ordering O_j is challenging. The difficulty lies in determining which category to assign a segment to which falls between any two categories in order to generate the minimum and maximum values for the tests. The segments of the circle between two neighboring observations among the categories are s_a, s_b, s_c, and s_d. Specifically, s_a represents the segment between categories C_{11} and C_{12}, s_b corresponds to the segment between C_{12} and C_{22}, s_c denotes the segment between C_{22} and C_{21}, and s_d uses the segment between C_{21} and C_{11}. Therefore, to obtain the maximum and minimum values for independence test statistics, we need to assign the segment values between two categories in a way that minimises and maximises the resulting test statistics. All future observations within one segment can be entirely assigned to one of the two categories at a time, or they can be shared between two categories. For example, if the segment s_a between C_{11} and C_{12} has a value of 3, the possible combinations are (3,0), (0,3), (1,2), and (2,1). Similar considerations apply to s_b, s_c, and s_d. We then apply the independence test on the combinations of segments and select the minimum and maximum values of the test statistics. However, determining the minimum and maximum values for some tests of independence such as the chi-square test is challenging. Therefore, an algorithm is needed to generate all combinations, calculate the minimum and maximum values, and perform the computation of the lower and upper RP. Algorithm 3 calculates NPI-RP approximations for tests of independence.

Algorithm 3 provides a systematic approach to approximate the NPI lower and upper reproducibility probabilities for tests of independence by sampling orderings of future observations among the data observations. This method is particularly useful when dealing with large datasets where considering all possible orderings is computationally infeasible. The algorithm begins by applying the test of independence, such as the chi-square or likelihood ratio test, to the original sample to make a decision about the null hypothesis H_0. The outcome of this test is recorded as C^*, where $C^* = 1$ if H_0 is rejected and $C^* = 0$ if it is not rejected.

In the next step, a specific ordering of the n future observations among the corresponding n data observations is sampled, denoted by the vector $(s_1, s_2, \ldots, s_n)$.

Algorithm 3 NPI-RP approximations for test of independence

```
1: Apply the test of independence on the original sample, and make a decision about $H_0$, then record the test outcome: $C^* = 1$ if $H_0$ is rejected and $C^* = 0$ if $H_0$ is not rejected.
2: Sample a specific ordering of the $n$ future observations among the corresponding $n$ data observations $(s_1, s_2, \dots, s_n)$.
3: For this ordering in Steps 2, calculate possible combinations of segments between two categories using $t_1, t_2, t_3, t_4$:
       $n_{11} = s^{11} + t_1 + t_2, \qquad t_1 \in \{0, 1, \dots, s_a\},$
       $n_{12} = s^{12} + (s_a - t_1) + t_3, \qquad t_2 \in \{0, 1, \dots, s_d\},$
       $n_{21} = s^{21} + (s_d - t_2) + t_4, \qquad t_3 \in \{0, 1, \dots, s_b\},$
       $n_{22} = s^{22} + (s_b - t_3) + (s_c - t_4), \qquad t_4 \in \{0, 1, \dots, s_c\},$
4: Compute the $\chi^2$ or $G^2$ test statistics and corresponding p-value for all combinations from Steps 3 to find the minimum and maximum of the $\chi^2$ or $G^2$.
5: Perform Steps 2-4 $n^*$ times to obtain $n^*$ values of the minimum and maximum $\chi^2$ or $G^2$, and each time record the test outcome : $C^*_{\text{Min}_j} = 1$ , $C^*_{\text{Max}_j} = 1$, if $H_0$ is rejected, and $C^*_{\text{Min}_j} = 0$, $C^*_{\text{Max}_j} = 0$ if $H_0$ is not rejected.
6: Approximate the NPI lower and upper probabilities for test reproducibility of the $n^*$ sampled ordering, for the case that $H_0$ was rejected for the original test data:
       $\underline{RP} = \frac{1}{n^*} \sum_{j=1}^{n^*} \mathbb{1}(C^*_{\text{Min}_j} = 1)$
       $\overline{RP} = \frac{1}{n^*} \sum_{j=1}^{n^*} \mathbb{1}(C^*_{\text{Max}_j} = 1)$
7: Approximate the NPI lower and upper probabilities for test reproducibility of the $n^*$ sampled ordering, for the case that $H_0$ was not rejected for the original test data:
       $\underline{RP} = \frac{1}{n^*} \sum_{j=1}^{n^*} \mathbb{1}(C^*_{\text{Min}_j} = 0)$
       $\overline{RP} = \frac{1}{n^*} \sum_{j=1}^{n^*} \mathbb{1}(C^*_{\text{Max}_j} = 0)$
```

For this sampled ordering, we calculate all possible combinations of future observations allocated to the segments between neighboring categories on the circle. This involves using variables t_1, t_2, t_3, and t_4, where each t_k ranges from 0 to the total number of future observations in the corresponding segment. These variables represent the number of future observations assigned to each segment between the categories. We then compute the counts for the contingency table based on these allocations. By exploring all possible values of t_k, we generate all combinations of future observations in these segments.

After calculating the counts for each combination, we compute the test statistics (either χ^2 or G^2) and the corresponding p-values. This allows us to identify the minimum and maximum values of the test statistics for the sampled ordering. We repeat this process n^* times, each time sampling a new ordering of future observations and calculating the associated minimum and maximum test statistics. We record whether the null hypothesis H_0 would be rejected at the minimum and maximum test statistics, denoted as $C^*_{\text{Min}_j}$ and $C^*_{\text{Max}_j}$, respectively. Specifically, $C^*_{\text{Min}_j} = 1$ if H_0 is rejected at the minimum test statistic and $C^*_{\text{Min}_j} = 0$ otherwise; similarly for $C^*_{\text{Max}_j}$.

Finally, we approximate the NPI lower and upper reproducibility probabilities based on these outcomes. If the original test rejected H_0, the NPI lower reproducibility probability is approximated by the proportion of sampled orderings where H_0 is also rejected at the minimum test statistic. The NPI upper reproducibility probability is approximated by the proportion where H_0 is rejected at the maximum test statistic. Conversely, if the original test did not reject H_0, we calculate the lower and upper reproducibility probabilities based on the proportion of times H_0 is not rejected in the sampled orderings.

4.2.2 NPI-B-RP for tests of independence

In the previous section, NPI sampling of orderings was introduced to study the NPI-RP for chi-square test of independence and the likelihood ratio test of independence. However, NPI sampling of orderings may not always be feasible because deriving the exact NPI lower and upper reproducibility probabilities can be challenging for some test statistics. In this section, we introduce the NPI-B-RP method to approximate the reproducibility probability for chi-square test of independence and the likelihood ratio test of independence. The NPI-B-RP method uses a single point estimate to represent the NPI reproducibility probability rather than providing lower and upper reproducibility probabilities. We use NPI-Bootstrap, as described in Section 3.5, to study the reproducibility of the chi-square and likelihood ratio tests of independence.

The NPI-B-RP for the test of independence operates as follows. First, the test of independence is applied to the original sample, and the outcome is recorded as $C^* = 1$ if the null hypothesis H_0 is rejected, and $C^* = 0$ if H_0 is not rejected. Next, an NPI-B sample is drawn based on the original sample, and the test of independence is applied to this sample to obtain a test result. This step is repeated B times, where each repetition is indexed by $j = 1, \ldots, B$. For each test result, the decision is recorded as $C_j^* = 1$ if H_0 is rejected, and $C_j^* = 0$ if H_0 is not rejected. Following this, the relative proportion rp is calculated as $rp = \sum_{j=1}^{B} \mathbb{1}(C^* = C_j^*)/B$, where $\mathbb{1}$ is an indicator function. Finally, steps 2 through 4 are repeated a total of h times, resulting in a set of rp values denoted by $rp_1, rp_2, \ldots, rp_h$. Algorithm 4 applies the NPI-B approach to approximate the reproducibility probability for the chi-square and likelihood ratio tests of independence. To achieve reliable results, it is required to use larger values of B, such as 1000 or 2000 replications [33]. In Algorithm 4, the number of runs $h = 100$ and the number of bootstrapped samples per run $B = 1000$.

4.2.3 Examples

This section studies the reproducibility probability using chi-square and likelihood ratio tests of independence. In Example 4.2.1, simulated data is used to investigate reproducibility probability with the NPI-RP and NPI-B-RPapproaches, and the results are then compared. In Example 4.2.2, data sets from the literature are analysed to investigate reproducibility probability using the NPI-RP and NPI-B-RP approaches. The results are compared to evaluate the effectiveness of these methods

Algorithm 4 NPI-B-RP for the test of independence

```
Apply the test of independence on the original sample, then record the test outcome: $C^* = 1$ if $H_0$ is rejected and $C^* = 0$ if $H_0$ is not rejected.
Draw a NPI-B sample based on the original sample and apply the test of independence and obtain the results.
Repeat step 2 $B$ times for $j = 1, \dots, B$ and each time record the test decision: $C_j^* = 1$ if $H_0$ is rejected and $C_j^* = 0$ if $H_0$ is not rejected..
Calculate rp, where $rp = \sum_{j=1}^{B} \mathbb{1}(C^* = C_j^*)/B$.
Perform Steps 2-4 in total $h$ times, leading to $rp$ by $rp_1, rp_2, \cdots, rp_h$.
```

	Control	Treatment	Total
Male	14	5	19
Female	6	15	27
Total	20	20	40

Table 4.1: Distribution of participants in control and treatment groups based on gender.

on a large dataset.

Example 4.2.1.

We explore NPI-RP for both the chi-square test and the likelihood ratio test of independence for the dataset in Table 4.1. The chi-square test statistic is $\chi^2 = 8.1203$, with a corresponding p-value of 0.0044. The critical value is 3.84 at a significance level of 0.05 with one degree of freedom, so we reject the null hypothesis (H_0). This outcome indicates a statistically significant relationship between gender and treatment.

Table 4.2 presents the approximated NPI lower and upper reproducibility probabilities and the corresponding 95% confidence intervals for various numbers of sampled orderings n^* for the chi-square test of independence. These confidence intervals are calculated using the standard normal approximation outlined in Equation (2.7.3).

The likelihood ratio test statistic is $G^2 = 8.4238$, and with a significance level of 5% and one degree of freedom, the null hypothesis (H_0) is again rejected. Table 4.3 presents the approximated NPI lower and upper reproducibility probabilities and the corresponding 95% confidence intervals for different values of n^* for the likelihood ratio test of independence.

In this example, the likelihood ratio test and chi-square test of independence produce similar results in assessing reproducibility. As shown in Tables 4.2 and 4.3, reasonably accurate approximations of the $\underline{RP}$ and $\overline{RP}$ can be achieved when the number of sampled orderings is equal to or greater than 2,000. According to Coolen and Marques [24], sampling 2,000 orderings typically provides a reliable impression of reproducibility for most practical applications. This quantity is relatively small compared to the total number of possible orderings. When sampling is increased to 100,000 orderings, it enables highly accurate approximations of the NPI lower and upper reproducibility probabilities, though calculations with this number of orderings require more computational time. NPI sampling of orderings offers a computationally efficient method for approximating the lower and upper reproducibility probabilities.

Using Algorithm 4 for Example 4.2.1, we obtained summary statistics including the minimum, mean, and maximum values from $rp_1, rp_2, \ldots, rp_{100}$. The mean of NPI-B-RP values for the chi-square and likelihood ratio tests of independence are 0.832 and 0.838, respectively. The mean of the NPI-B-RP values lies between the approximations of the $\underline{RP}$ and $\overline{RP}$ reproducibility probabilities of the chi-square and likelihood ratio tests of independence for $n^* = 100$ to $n^* = 100{,}000$.

Example 4.2.2. We explore NPI-RP for both the chi-square test and likelihood ratio test of independence, with a sample size of $n = 156$, as shown in Table 4.4. The data in Table 4.4 refer to the distance walked by 156 patients with degenerative lumbar stenosis with neurogenic intermittent claudication before and after surgery [4].

The chi-square test statistic is $\chi^2 = 12.18399$ and the corresponding p-value is 0.00048, The critical value is 3.84 at a significance level of 0.05 and one degree of freedom so the null hypothesis H_0 is rejected. By rejecting H_0, we conclude that

n^*	$\underline{RP}$	95% CI	$\overline{RP}$	95% CI
20	0.5000	(0.3319, 0.7680)	0.8000	(0.6246, 0.9753)
100	0.5800	(0.4832, 0.6767)	0.8700	(0.8040, 0.9359)
500	0.5860	(0.5420, 0.6290)	0.9140	(0.8894, 0.9385)
1000	0.5880	(0.5574, 0.6185)	0.8900	(0.8706, 0.9093)
2000	0.5820	(0.5603, 0.6036)	0.9005	(0.8873, 0.9136)
5000	0.5854	(0.5717, 0.5990)	0.9036	(0.8954, 0.9117)
10,000	0.5881	(0.5784, 0.5977)	0.8935	(0.8874, 0.8995)
20,000	0.5930	(0.5861, 0.5998)	0.8969	(0.8927, 0.9011)
50,000	0.5896	(0.5852, 0.5939)	0.8960	(0.8933, 0.8987)
100,000	0.5898	(0.5867, 0.5928)	0.8959	(0.8940, 0.8978)

Table 4.2: $\underline{RP}$ and $\overline{RP}$ for an observed sample of Example 4.2.1 and for χ^2 test for different values of n^*.

there is a statistically significant association between the walking distance before surgery and walking distance after surgery. Table 4.5 presents the approximated NPI lower and upper reproducibility probabilities and the corresponding 95% confidence intervals with different numbers of orderings sampled n^* for the chi-square test of independence. The likelihood ratio test statistic is $G^2 = 12.40068$ and p-value is 0.00042, the H_0 is reject as sign level 5%. Table 4.6 presents the approximated NPI lower and upper reproducibility probabilities and the corresponding 95% confidence intervals with different numbers of orderings sampled n^* for likelihood ratio test of independence. A comparison of the reproducibility of likelihood ratio test of independence and chi-square test of independence indicated that their results were closely similar. As n increases, the difference between $\underline{RP}$ and $\overline{RP}$ becomes smaller, reflecting the large amount of information. Additionally, as we increase the number of sampled orderings, we observe that the confidence interval for the lower and upper reproducibility values becomes narrower

Additionally, we applied Algorithm 4 to Example 4.2.2. The mean of NPI-B-RP values for the chi-square and likelihood ratio tests of independence are 0.906 and 0.904, respectively. Similarly, the mean of NPI-B-RP value falls between the

n^*	$\underline{RP}$	95% CI	$\overline{RP}$	CI(0.95
20	0.6000	(0.3853, 0.8147)	0.8000	(0.6935, 1.0000)
100	0.6200	(0.5249, 0.7151)	0.9200	(0.8668, 0.9732)
500	0.5960	(0.5529, 0.6390)	0.8960	(0.8692, 0.9227)
1000	0.6070	(0.5767, 0.6373)	0.9050	(0.8868, 0.9232)
2000	0.5745	(0.5528, 0.5962)	0.9010	(0.8879, 0.9141)
5000	0.5880	(0.5744, 0.6016)	0.8880	(0.8793, 0.8967)
10,000	0.5954	(0.5858, 0.6050)	0.8984	(0.8925, 0.9043)
20,000	0.6005	(0.5937, 0.6073)	0.9011	(0.8970, 0.9052)
50,000	0.5953	(0.5910, 0.5996)	0.8973	(0.8947, 0.9000)
100,000	0.5929	(0.5899, 0.5959)	0.8998	(0.8979, 0.9017)

Table 4.3: $\underline{RP}$ and $\overline{RP}$ for an observed sample of for the data set in Example 4.2.1 and for G^2 test for different values of n^*.

approximations of the $\underline{RP}$ and $\overline{RP}$ reproducibility probabilities of the chi-square and likelihood ratio tests of independence for $n^* = 100$ to $n^* = 100,000$.

4.2.4 Simulation studies

In this study, we investigate NPI-B-RP and approximate NPI-RP for the chi-square test of independence and the likelihood ratio test of independence. We generate data from a multinomial distribution with probabilities $(0.25, 0.25, 0.25, 0.25)$ under the null hypothesis and with probabilities $(0.1, 0.5, 0.2, 0.2)$ under the alternative hypothesis. The simulations were conducted by sampling under the alternative hypothesis, as it resulted in more test statistics being close to the test threshold. This simulation study uses the following inputs: $n = 35, 70$, and $N = 50$ simulations. We examine the effect of increasing sample size on the patterns of RP values using simulations. The critical value is 3.84 at a significance level of 0.05 and one degree of freedom. For a sample size of $n = 35$, the power of the test with a medium effect size $w = 0.3$ is approximately 0.426, and with a large effect size $w = 0.5$, the power is approximately 0.84. For a sample size of $n = 70$, the power of the test

		After Surgery		Total
		$\leq$ 500m	$>$ 500m	
Before Surgery	$\leq$ 500m	56	37	93
	$>$ 500m	20	43	63
Total		76	80	156

Table 4.4: Walking distance before and after surgery in 156 patients with lumbar stenosis.

n^*	$\underline{RP}$	95% CI	$\overline{RP}$	95% CI
100	0.8700	(0.8040,0.9359)	0.9500	(0.9072, 0.9927)
500	0.8260	(0.7927,0.8592)	0.9140	(0.8894, 0.9385)
1000	0.8250	(0.8014, 0.8485)	0.9070	(0.8889, 0.9250)
2000	0.8160	(0.7990, 0.8329)	0.9150	(0.9027, 0.9272)
5000	0.8158	(0.8050, 0.8265)	0.9164	(0.9087, 0.9240)
10,000	0.8109	(0.8032, 0.8185)	0.9160	(0.9105, 0.9214)
20,000	0.8147	(0.8093, 0.8201)	0.9157	(0.9119, 0.9195)
50,000	0.8123	(0.8088, 0.8157)	0.9159	(0.9134, 0.9183)
100,000	0.8155	(0.8130, 0.8178)	0.9167	(0.9149, 0.9183)

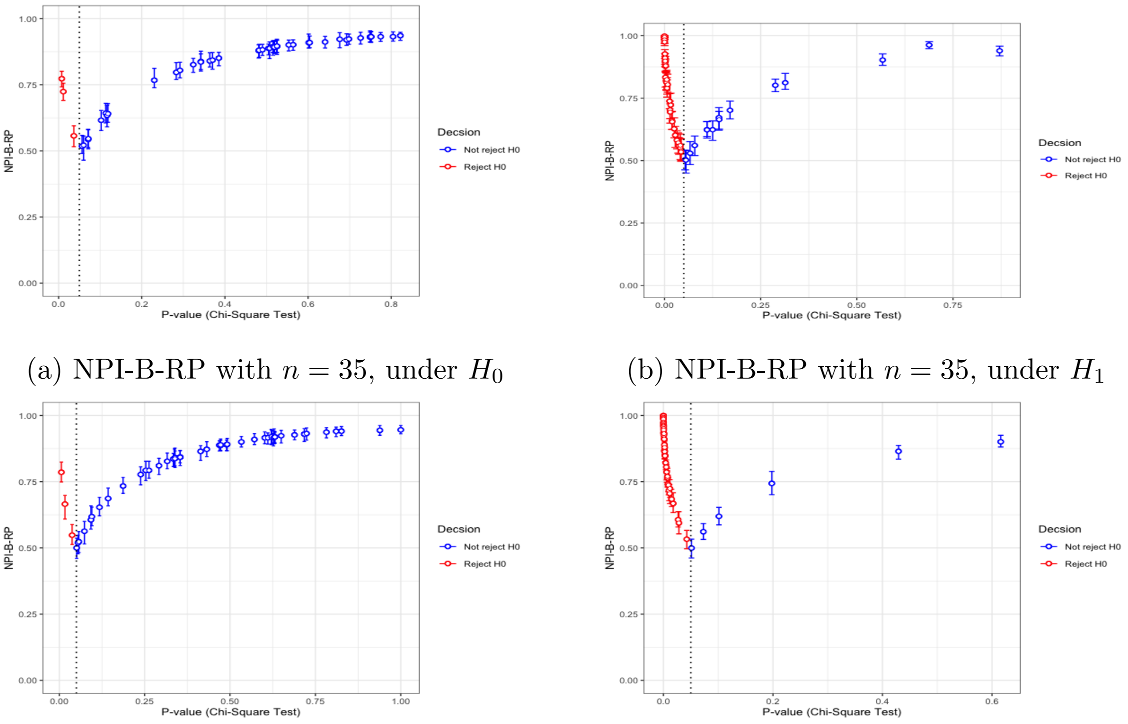
Table 4.5: $\underline{RP}$ and $\overline{RP}$ for an observed sample for the data set in Example 4.2.2 and and for χ^2 test for different values of n^*.

with a medium effect size $w = 0.3$ is approximately 0.70, and with a large effect size $w = 0.5$, the power is approximately 0.98. The NPI-B method was used to determine all RP values using $B = 1000$ bootstrap samples and $h = 100$. A sample of size n is generated for each run from each of these multinomial distributions, the chi-square test of independence is performed, and the corresponding p-value is calculated. The NPI-B-RP and $\underline{RP}$ and $\overline{RP}$ are approximated for the chi-square test of independence. The RP value is calculated using the NPI-B-RP and NPI-RP methods, as outlined in Algorithms 3 and 4. Plots of these metrics for two different sample sizes are presented in Figures 4.1 and 4.2, where simulations are conducted under H_0 and H_1.

n^*	$\underline{RP}$	95% CI	$\overline{RP}$	95% CI
100	0.8000	(0.7216, 0.8783)	0.9300	(0.8799, 0.9800)
500	0.8180	(0.7841, 0.8518)	0.9340	(0.9122, 0.9557)
1000	0.8140	(0.7898, 0.8381)	0.9100	(0.8922, 0.9277)
2000	0.8155	(0.7985, 0.8324)	0.9135	(0.9011, 0.9258)
5000	0.8152	(0.8044, 0.8259)	0.9178	(0.9101, 0.9254)
10,000	0.8133	(0.8056, 0.8209)	0.9175	(0.9121, 0.9228)
20,000	0.8163	(0.8109, 0.8216)	0.9152	(0.9113, 0.9190)
50,000	0.8124	(0.8089, 0.8158)	0.9140	(0.9115, 0.9164)
100,000	0.8136	(0.8112, 0.8160)	0.9172	(0.9154, 0.9188)

Table 4.6: $\underline{RP}$ and $\overline{RP}$ for an observed sample for the data set in Example 4.2.2 and for G^2 test for different values of n^*.

In this simulations, we examine the relationship between NPI-B-RP with the p-value for the chi-square test of independence. The minimum, mean and maximum are taken from $rp_1, rp_2, \cdots, rp_{100}$ from the Algorithm 4. Figure 4.1 shows RP values using the NPI-Bootstrap method under H_0 and H_1 for samples of size 35 and 70. In general, the RP increases when the p-value moves away from the threshold of 0.05. Based on the figures, it is evident that reproducibility is lowest around when the observed p-value is close to the threshold 0.05. When the original test statistic is close to the test threshold, NPI-B-RP is approximately 0.5. As long as further information is not available, a repeat experiment would result in a second test statistic that is equally likely to be larger or smaller than the original test statistic, and therefore would result in a probability of 0.5 for the same conclusion. Repeating an experiment with a test statistic that is far from the test threshold will likely produce a second test statistic that is far from the test threshold as well. Consequently, when the test statistic moves away from the test thresholds, the RP values tend to increase. When the p-value is 0.25, the NPI-B-RP is 0.75 or higher in non-rejection cases. There are similar patterns observed in applications of NPI reproducibility for different test scenarios [11, 67].



(a) NPI-B-RP with $n = 35$, under H_0

(b) NPI-B-RP with $n = 35$, under H_1

(c) NPI-B-RP with $n = 70$, under H_0

(d) NPI-B-RP with $n = 70$, under H_1

Figure 4.1: Simulations under H_0 and H_1: NPI-B-RP values for chi-square test of independence.

As the sample size increases, both rejection and non-rejection approaches 0.5 when p-values are close to the threshold 0.05, as well as RP values become less variable. This occurs because increased sample size leads to an increase in the power of tests. The patterns of RP change when simulations are conducted under alternative hypothesis due to the change in p-value with respect to the the threshold 0.05. For the simulations under H_1, increasing the sample size leads to more rejection cases because the test becomes more powerful with a larger sample size.

In these simulations, the reproducibility is examined for the chi-square of independence using the NPI sampling of orderings method. The effect of the number of sampled orderings on the lower and upper RP values is examined using $n^* = 2000$ and $n^* = 5000$. We approximate NPI-RP by sampling of orderings rather than considering all possible orders. Figures 4.2 show RP values using NPI sampling of orderings method under H_0 and H_1 for samples of size 35 and 70. Figures 4.2 indicate that there are no substantial differences on the patterns for different values of

n^*. The number of orderings sampled can be increased in order to obtain more accurate approximations of the $\underline{RP}$ and $\overline{RP}$. When the sample size is small, the lower reproducibility probabilities tend to approach 0.25 when the observed test statistics are near the 0.05 threshold. In both sample sizes, the $\underline{RP}$ and $\overline{RP}$ tend to be less than 0.75 when p-values are close to the threshold 0.05. When n increases from 35 to 70, the $\underline{RP}$ value tends to increase when the observed p-value of the chi-square test approaches the threshold 0.05 and the difference between the lower and upper RPs becomes narrower. There is a positive relationship between the power of the test and the sample size, which means that a larger sample size gives a greater power.

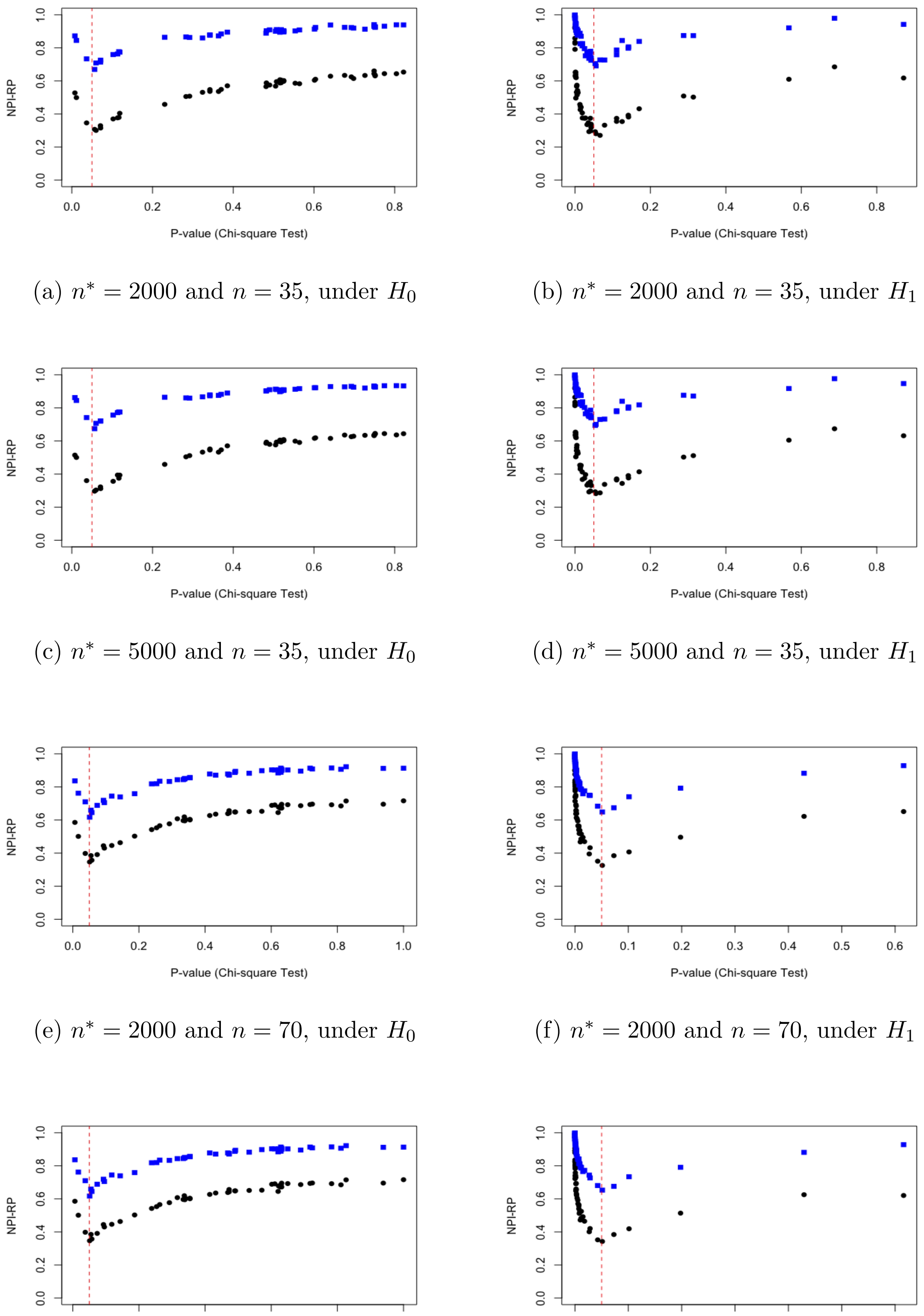
Figure 4.2 focuses on approximated upper and lower probabilities, which represent imprecise probabilities. In contrast, Figure 4.1 illustrates the Bootstrap-RP method, providing a point estimate of the reproducibility probability (RP) instead of the interval defined by the lower and upper NPI-RP values. The Bootstrap-RP method is repeated 100 times, generating values $rp_1, rp_2, \cdots, rp_{100}$. From these values, the minimum and maximum are identified to determine the range for Bootstrap-RP. Throughout the thesis, red circles indicate RP values for cases where the null hypothesis is rejected, while blue circles represent RP values for cases where the null hypothesis is not rejected.

The NPI-B-RP and NPI-RP are compared for the chi-square test with NPI-B-RP and NPI for the likelihood ratio test. Note that the simulated data are the same for chi-square test of independence and the likelihood ratio test of independence. Figures 4.3 and 4.4 show RP values for the likelihood ratio test of independence with two different sample sizes. The likelihood ratio test of independence produces similar results as the chi-square test of independence. There were very similar results when comparing the reproducibility of these two tests with simulated data under H_0 and H_1. The likelihood ratio test and chi-square test are closely related in many applications, particularly in large sample settings. As a result, it is expected that both tests would yield similar RP values. Additional simulation results for NPI-B-RP and NPI-RP of the chi-square test and the likelihood ratio test are provided in Appendix A. We tested various values of n and observed a similar impact each time, so we limited it to two values. This indicates that the RP values exhibit less fluctuation as the sample size increases, as clearly shown in the figures presenting the simulation results.

4.3 Reproducibility of McNemar's test

In this section, the reproducibility of McNemar's test is investigated using the NPI sampling of orderings and NPI-B methods. The null hypothesis is $H_0 : \pi_{12} = \pi_{21}$ and the alternative hypothesis is $H_1 : \pi_{12} \neq \pi_{21}$, the level of significance is 0.05. In this simulation study, the following inputs are used: $n = 35, 70$, and $N = 50$ simulations per run. Simulations are conducted to evaluate the NPI and NPI-B methods for the RP of McNemar test of RP in accordance with the same steps as for the test of independence described in Section 4.2. The simulations are performed both under H_0 and H_1. Under H_0, data are generated from the a multinomial distribution with probabilities (0.25, 0.25, 0.25, 0.25). Under H_1 data are generated from a multinomial distribution with probabilities $(0.1, 0.5, 0.2, 0.2)$. The power of McNemar's test with cell probabilities $(0.1, 0.5, 0.2, 0.2)$ under the alternative hypothesis is approximately 0.566 for a sample size of $n = 35$ and approximately 0.851 for a sample size of $n = 70$. The NPI-B-RP, $\underline{RP}$, and $\overline{RP}$ are provided for the McNemar's test, and plots of these metrics for two different sample sizes under H_0 and H_1 are presented in Figures 4.5 and 4.6.

We examine the relationship between NPI-B-RP with the p-value for the McNemar's test. The minimum, mean and maximum of rp by $rp_1, rp_2, \cdots, rp_{100}$ are computed. Figure 4.5 displays RP values using the NPI-Bootstrap method under H_0 and H_1 for samples of size 35 and 70. The RP tends to increase when the p-value is moved away from the the threshold 0.05. As can be seen from the figures, reproducibility is at its lowest around the threshold of the test. When the p-value is 0.75 the NPI-B-RP is higher than 0.85 in non-rejection cases (blue cases). In the figures, NPI-B-RP tends to be lower when the null hypothesis is rejected (red cases) than when it is not rejected (blue cases). When sample sizes increase, both rejection and non-rejection RP approach 0.5 when p-values are close to the the threshold 0.05, and RP values become less variable. This occurs because increased sample



(a) $n^* = 2000$ and $n = 35$, under H_0

(b) $n^* = 2000$ and $n = 35$, under H_1

(c) $n^* = 5000$ and $n = 35$, under H_0

(d) $n^* = 5000$ and $n = 35$, under H_1

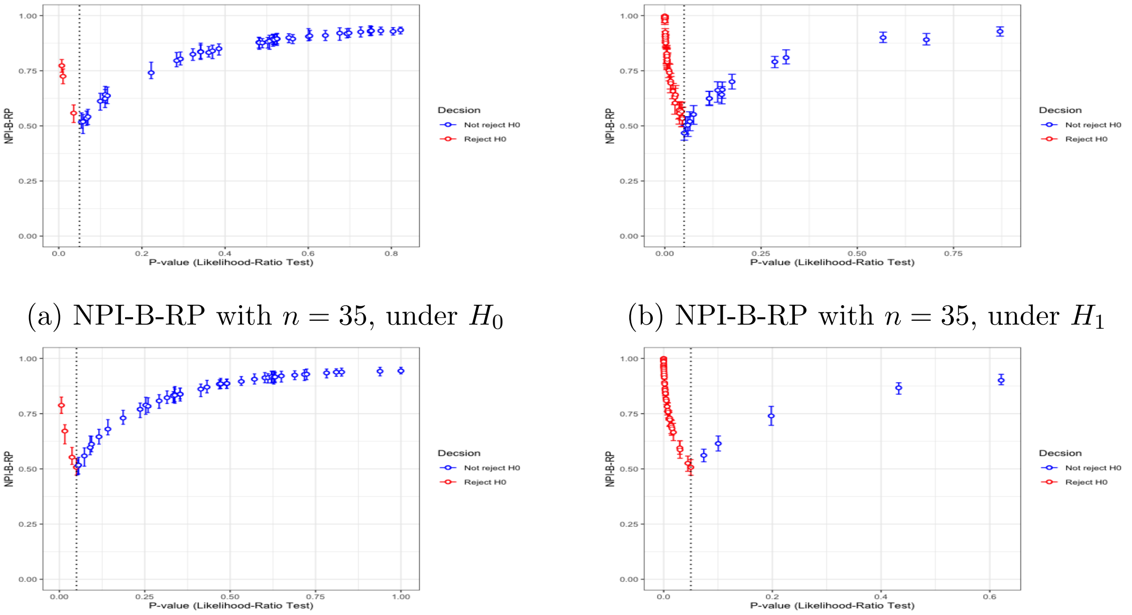
(e) $n^* = 2000$ and $n = 70$, under H_0

(f) $n^* = 2000$ and $n = 70$, under H_1

(g) $n^* = 5000$ and $n = 70$, under H_0

(h) $n^* = 5000$ and $n = 70$, under H_1

Figure 4.2: Approximation values of the upper (blue) and lower (black) RPs for 50 replications, for chi-square testing of independence.



(a) NPI-B-RP with $n = 35$, under H_0

(b) NPI-B-RP with $n = 35$, under H_1

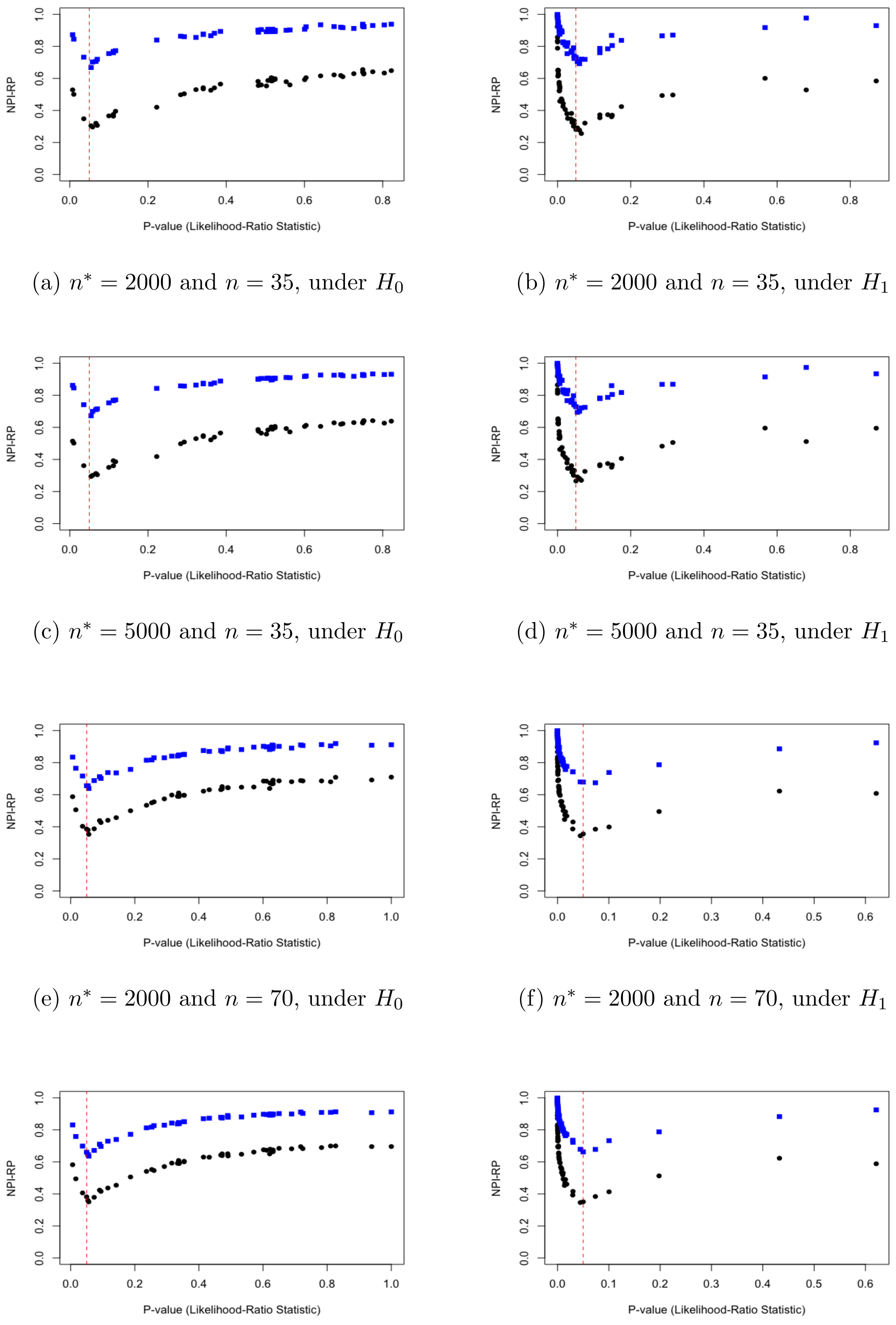
(c) NPI-B-RP with $n = 70$, under H_0

(d) NPI-B-RP with $n = 70$, under H_1

Figure 4.3: Simulations under H_0 and H_1: NPI-B-RP values for likelihood ratio test of independence.

size leads to an increase in the power of the test. The patterns of NPI-B-RP change when simulations are conducted under alternative hypotheses due to the change in p-value with respect to the threshold 0.05. In the simulations under H_1, increasing the sample size results in more rejection cases since a larger sample size makes the test more powerful.

The reproducibility of McNemar's test is examined using the NPI sampling of orderings method. Figure 4.6 shows RP values using the NPI sampling of orderings method under H_0 and H_1 for samples of size 35 and 70. The p-value of McNemar's test, $\overline{RP}$, and $\underline{RP}$ were determined for each simulated sample. The effect of the sample size on $\overline{RP}$ and $\underline{RP}$ is examined using the number of orderings sampled, $n^* = 2000$ and $n^* = 5000$. In both sample sizes, the values of $\underline{RP}$ and $\overline{RP}$ tend to be less than 0.75 when the p-values are close to the threshold 0.05. Additionally, as the sample size increases, the $\underline{RP}$ value approaches 0.4, and the observed p-value of McNemar's test approaches the threshold 0.05. Both $\overline{RP}$ and $\underline{RP}$ tend to increase when the p-value moves away from the threshold 0.05. Furthermore, as the sample



(a) $n^* = 2000$ and $n = 35$, under H_0

(b) $n^* = 2000$ and $n = 35$, under H_1

(c) $n^* = 5000$ and $n = 35$, under H_0

(d) $n^* = 5000$ and $n = 35$, under H_1

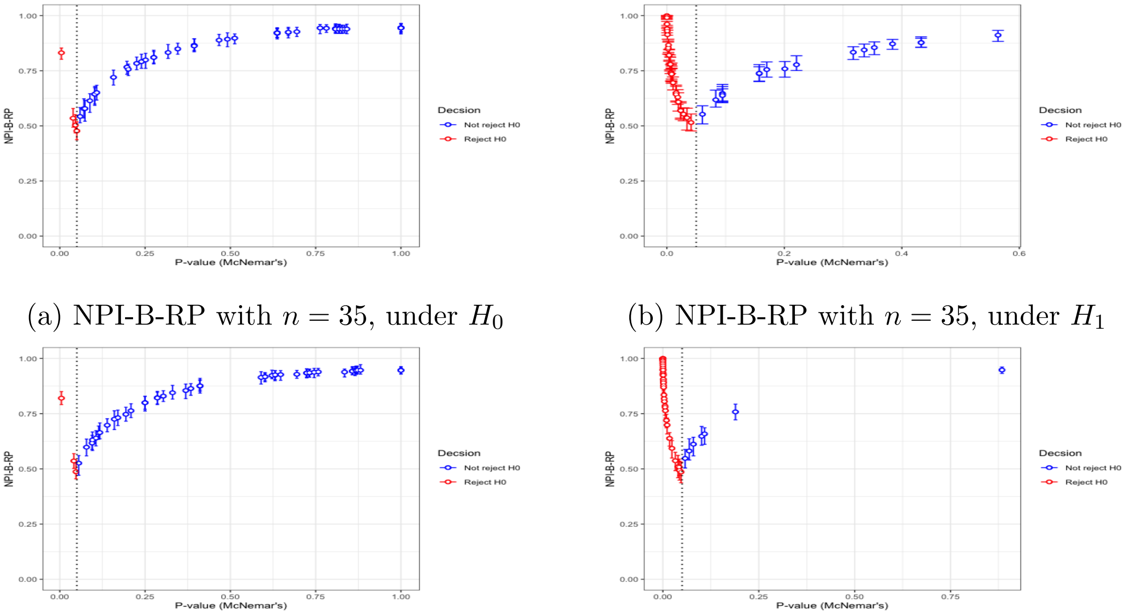
(e) $n^* = 2000$ and $n = 70$, under H_0

(f) $n^* = 2000$ and $n = 70$, under H_1

(g) $n^* = 5000$ and $n = 70$, under H_0

(h) $n^* = 5000$ and $n = 70$, under H_1

Figure 4.4: Approximation values of the upper (blue) and lower (black) RPs for 50 replications, for likelihood ratio testing of independence.



(a) NPI-B-RP with $n = 35$, under H_0

(b) NPI-B-RP with $n = 35$, under H_1

(c) NPI-B-RP with $n = 70$, under H_0

(d) NPI-B-RP with $n = 70$, under H_1

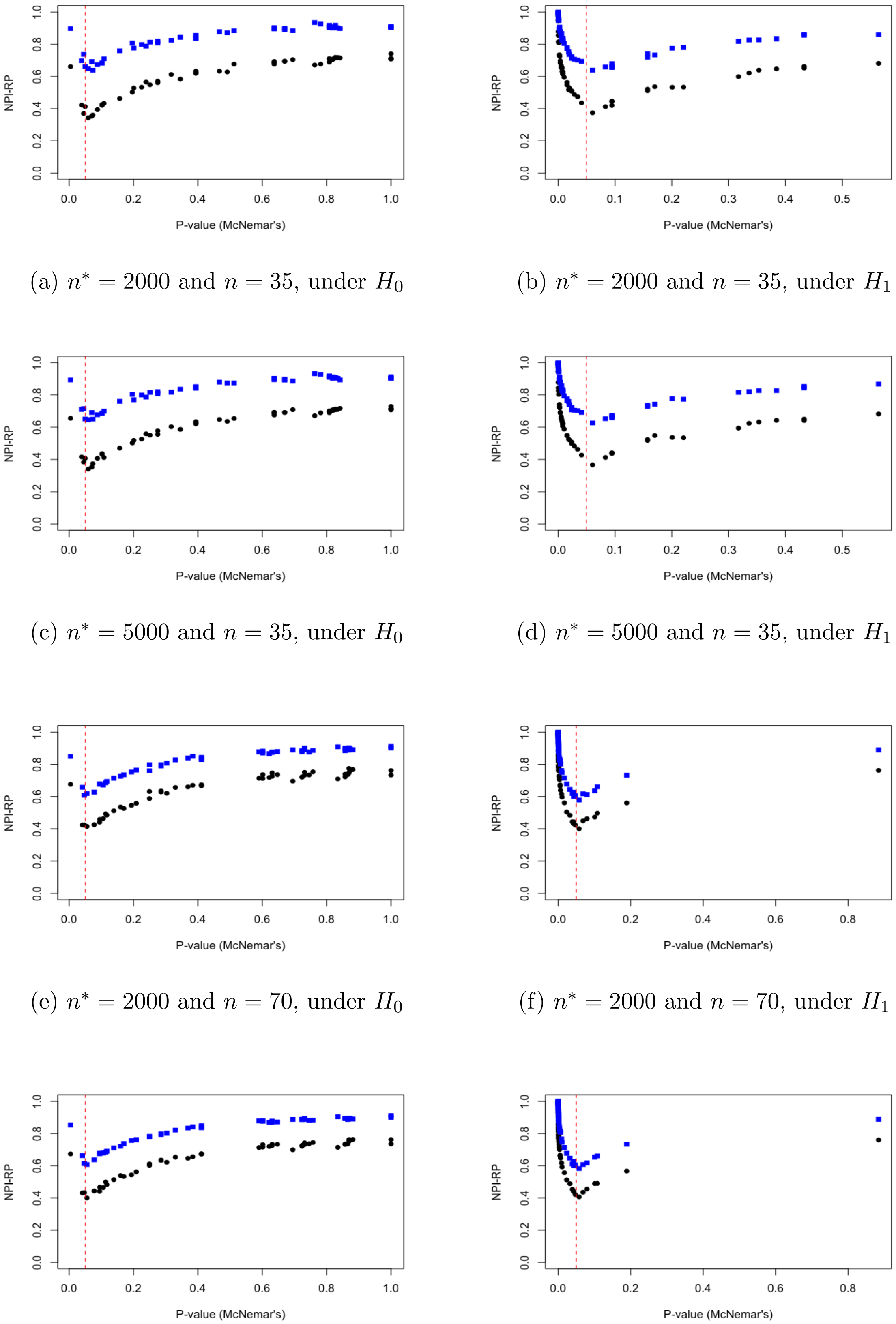
Figure 4.5: Simulations under H_0 and H_1: NPI-B-RP for values McNemar's test.

size increases, the difference between the lower and upper RPs becomes smaller, reflecting the large amount of information. The power of the test is positively related to sample size, meaning that a larger sample size gives greater power. Additional simulation results for NPI-B-RP and NPI-RP of the McNemar's test are provided in Appendix A.

We compared the reproducibility of the chi-square test of independence, the likelihood ratio test of independence, and McNemar's test using the sampling of orderings and NPI-B methods. Note that all simulated data sets are the same for all the tests. McNemar's test produces similar results as the chi-square test of independence and the likelihood ratio test of independence for reproducibility for both sampling of orderings and NPI-B methods.

4.4 Reproducibility of Fisher's exact test

In this section, the reproducibility for the Fisher's exact test of independence is studied using the NPI sampling of orderings and NPI-B methods. The Fisher's exact test is a powerful statistical method commonly used when there are relatively small



(a) $n^* = 2000$ and $n = 35$, under H_0

(b) $n^* = 2000$ and $n = 35$, under H_1

(c) $n^* = 5000$ and $n = 35$, under H_0

(d) $n^* = 5000$ and $n = 35$, under H_1

(e) $n^* = 2000$ and $n = 70$, under H_0

(f) $n^* = 2000$ and $n = 70$, under H_1

(g) $n^* = 5000$ and $n = 70$, under H_0

(h) $n^* = 5000$ and $n = 70$, under H_1

Figure 4.6: Approximation values of the upper (blue) and lower (black) RPs for 50 replications, for McNemar's test.

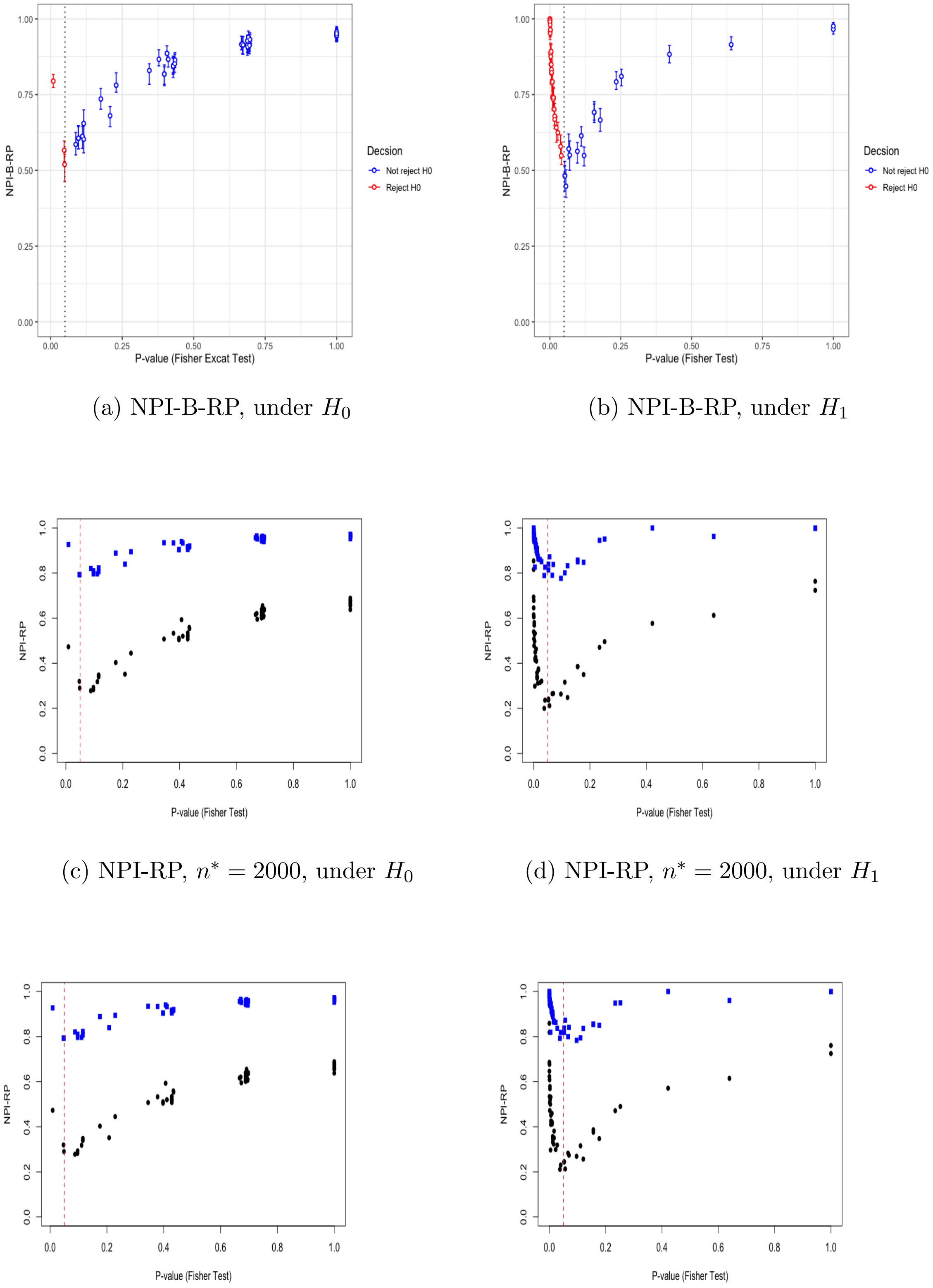
sample sizes. The data are generated under H_0 from a multinomial distribution with probabilities (0.25, 0.25, 0.25, 0.25), whereas under the under H_1 the data are generated from the multinomial distribution with probabilities (0.6, 0.1, 0.2, 0.2). In this simulation study, the following inputs are used: $n = 12$, and $N = 50$ simulations per run. Using the same steps as for the tests of independence in Section 4.2, simulated experiments are performed to evaluate the NPI sampling of orderings and NPI-B methods for the Fisher's exact test of RP. Plots of these metrics for sample size are displayed in Figure 4.7 under H_0 and H_1.

Based on the simulations, we investigate the performance of NPI-B-RP for Fisher's Exact test. Using $rp_1, rp_2, \cdots, rp_{100}$, the minimal, mean, and maximal values are calculated. Generally, the RP increases as the p-value moves away from the test threshold. Reproducibility appears to be lowest around the threshold 0.05, as shown in the figures. When the p-value is 0.25, the NPI-B-RP is higher then 0.75 in cases where the null hypothesis is not rejected (blue cases).It is clear that, as expected, the simulations indicate that RP values based on NPI-B show variability due to the small sample size. Similar findings with small sample sizes have been observed in previous NPI studies of test reproducibility [11, 67].

In this study, a Fisher's exact test is conducted on the reproducibility using the NPI sampling of orderings method. Each simulated sample was given a p-value of Fisher's exact test, and the $\overline{RP}$ and $\underline{RP}$ were calculated. The $\overline{RP}$ tends to approach 0.80 when p-values are close to thethreshold 0.05. As the observed p-value of a Fisher's exact test approaches the threshold 0.05, the $\underline{RP}$ value approaches 0.3. When the p-value moves away from the threshold 0.05, both $\overline{RP}$ and $\underline{RP}$ tend to increase.

4.5 Concluding remarks

Test reproducibility is well aligned with the explicit predictive nature of NPI. Test reproducibility is the probability that the same test outcome would be obtained if a test were repeated under identical circumstances with the same sample size. In this chapter, we explored the estimation of reproducibility for tests of independence,



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure 4.7: Simulations under H_0 and H_1: values of NPI-RP and NPI-B-RP for Fisher's exact test, where $n = 12$.

Fisher's exact test, and McNemar's test using NPI bootstrap and NPI sampling of orderings. Through simulation studies, the reproducibility of tests has been studied using the NPI ampling of orderings and NPI-B methods. By increasing sample size, NPI-B-RP become close to 0.5 in both rejection and non-rejection cases when the observed p-values approach the threshold 0.05. The variability in the RP values decreases when the sample size is increased which results in a more powerful test. The bootstrap approach to predicting RP avoids the difficulty of determining the lower and upper boundaries in NPI-RP. In NPI-B-RP, we present the RP as a point estimate rather than the lower and upper values. Reproducibility of chi-square test of independence, the likelihood ratio test of independence and McNemar's test were also compared in the simulations and similar results were obtained.

Chapter 5

Bayesian Inference for Reproducibility of Tests Based on 2×2 Tables

5.1 Introduction

Chapter 4 introduced the NPI for the reproducibility probability in hypothesis tests based on the 2×2 table. In this chapter, we use Bayesian inference to evaluate the reproducibility of hypothesis tests involving 2×2 contingency tables. Specifically, we apply a Bayesian approach to analyze the observed data by defining prior distributions and likelihood functions, employing a Dirichlet prior alongside a multinomial likelihood to derive the posterior distributions of the cell probabilities.

It is important to emphasize that while Bayesian inference is used to derive the posterior distributions of the cell probabilities, the hypothesis testing itself is not a Bayesian procedure. The focus of this chapter is on using predictive inference within the Bayesian framework to measure reproducibility. This involves assessing how well the posterior predictive distribution predicts future observations, providing insights into the reproducibility of statistical tests.

Billheimer [10] discusses predictive inference within the Bayesian framework. Billheimer [10] argues that statistical modeling should predict observable quantities and events rather than make inferences through hypothesis testing or parameter estimation. Billheimer's view is that rather than concentrating on unobservable parameters, attention should be directed towards observable events.

Billheimer [10] suggests that while Bayesian statistics allow for predictive inference, the focus should be on predicting future observations rather than inferring parameters. Bayesian statistics are not explicitly designed to be predictive like NPI. Bayesian methods integrate prior information with observed data to update beliefs about parameters, while NPI only uses a few modeling assumptions to make predictions directly from data without prior knowledge. In this chapter, we will use the same Algorithm 4 as discussed in Chapter 4. However, instead of sampling from the NPI-B, we will sample from the posterior predictive distribution, as described in Algorithm 5.

This chapter start with some background of Bayesian inference for 2×2 contingency tables in Section 5.2. Section 5.3 presents the Bayesian reproducibility using chi-square test of independence, likelihood ratio test of independence, McNemars test and, Fisher's exact test. Finally, Section 5.4 provides some concluding remarks.

5.2 Bayesian inference for 2×2 contingency table

Bayesian inference combines prior information with data based on Bayes' rule [13, 71]. In this framework, the likelihood function is fundamental for updating the prior distribution via Bayes' theorem. The posterior distribution, considered the updated probability distribution for the unknown parameters of a statistical model, is obtained by combining both the prior distribution and the observed data. The Bayesian framework using the Bayes rule can be expressed as:

$$p(\pi|\mathbf{n}) = \frac{p(\mathbf{n}|\pi).p(\pi)}{\pi(\mathbf{n})} = \frac{p(\mathbf{n}|\pi).p(\pi)}{\int_\pi p(\mathbf{n}|\pi).p(\pi)d\pi} \tag{5.2.1}$$

where $p(\pi)$ is the prior distribution, $p(\mathbf{n}|\pi)$ is the likelihood function, and $p(\pi|\mathbf{n})$ is the posterior distribution and $\mathbf{n}$ is the observed data. Bayesian inference for 2×2 table uses a prior distribution on the parameters and expresses the results in the form of a posterior distribution.

Bayesian inference is very useful for analysing contingency tables. Early studies by Good [38, 39] and Lindley [54] showed the application of Bayesian methods to such data, utilising Dirichlet priors with multinomial distributions for efficient calculations. Good [39] used this framework to estimate multinomial probabilities, while Hoadley [43] showed that the method effectively generates posterior samples that accurately represent the multinomial distribution.

Later, researchers like Albert and Gupta [5], Lecoutre and Camilo [53] and Kateri and Agresti [50] also used Bayesian methods with Dirichlet priors to analyse tables, especially 2×2 tables. They found this approach helpful for making predictions and understanding the data.

Consider a 2×2 table with observed data represented as $\mathbf{n} = (n_{11}, n_{12}, n_{21}, n_{22})$ with cell probabilities $\boldsymbol{\pi} = (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})$. The likelihood function is given by:

$$p(\mathbf{n}|\boldsymbol{\pi}) = \left(\frac{n!}{\prod_{i,j} n_{ij}!} \right) \prod_{i,j} \pi_{ij}^{n_{ij}} \tag{5.2.2}$$

This likelihood function is used for a 2×2 table because it calculates the probability of seeing the specific counts based on the cell probabilities. It assumes that the data follows a multinomial distribution, where each cell count is independently determined based on the corresponding cell probability [2].

Dirichlet distributions are conjugate priors for multinomial distributions [59]. The Dirichlet distribution, denoted by $\mathrm{Dir}(\alpha_1, \ldots, \alpha_k)$, has parameters $\alpha_i > 0$ for $i = 1, \ldots, k$. These parameters are associated with the probabilities $\pi_1, \ldots, \pi_k$ and determine their expected values, where each $\pi_i \geq 0$ and $\sum_{i=1}^{k} \pi_i = 1$. The probability density function is given by:

$$p(\boldsymbol{\pi}) = \frac{\Gamma(B)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \pi_i^{\alpha_i - 1} \tag{5.2.3}$$

where

$$B = \sum_{i=1}^{k} \alpha_i$$

The joint distribution for a 2×2 table can be represented using the Dirichlet distribution with parameters $\alpha_{11}, \alpha_{12}, \alpha_{21}, \alpha_{22}$. The conjugacy of the Dirichlet prior to the multinomial likelihood ensures that the posterior distribution remains within the Dirichlet family, making Bayesian updating straightforward [59]. Conjugacy refers to the property that the prior and posterior distributions are in the same family when combined with the likelihood function. We use conjugate priors like the Dirichlet distribution to maintain analytical simplicity and make computations easier. While non-conjugate priors can provide more flexibility or include additional prior information, they often require numerical methods such as Markov Chain Monte Carlo (MCMC) for posterior inference, which can be computationally intensive [13, 71].

The posterior distribution for the Dirichlet distribution with a multinomial likelihood function is given by:

$$\begin{aligned}
p(\boldsymbol{\pi}|\mathbf{n}) &= \frac{p(\mathbf{n}|\boldsymbol{\pi})\, p(\boldsymbol{\pi})}{p(\mathbf{n})} \\
&\propto p(\mathbf{n}|\boldsymbol{\pi}) \times p(\boldsymbol{\pi}) \\
&= \left(\frac{n!}{\prod_{i,j} n_{ij}!} \prod_{i,j} \pi_{ij}^{n_{ij}} \right) \times \left(\frac{\Gamma(B)}{\prod_{i,j} \Gamma(\alpha_{ij})} \prod_{i,j} \pi_{ij}^{\alpha_{ij}-1} \right) \\
&\propto \left(\prod_{i,j} \pi_{ij}^{n_{ij}} \right) \times \left(\prod_{i,j} \pi_{ij}^{\alpha_{ij}-1} \right) \\
&= \prod_{i,j} \pi_{ij}^{n_{ij}+\alpha_{ij}-1} \\
&= \mathrm{Dir}\left(\boldsymbol{\pi} \,\middle|\, n_{11}+\alpha_{11},\ n_{12}+\alpha_{12},\ n_{21}+\alpha_{21},\ n_{22}+\alpha_{22} \right)
\end{aligned}$$

To compute the posterior predictive distribution for a single multinomial trial, we start by integrating the likelihood of a new observation over the posterior distribution of the parameters. In the case of a single trial, the likelihood $p(\hat{n}_{ij} \mid \boldsymbol{\pi})$ simplifies to the probability π_{ij} of observing an outcome in cell (i, j). This means the integral becomes $\int \pi_{ij}\, p(\boldsymbol{\pi} \mid \mathbf{n})\, d\boldsymbol{\pi}$. Using the properties of the Dirichlet distribution, this expected value is calculated as $\frac{\alpha_{ij} + n_{ij}}{\alpha_0 + n}$, where $\alpha_0 = \sum_{i,j} \alpha_{ij}$ is the sum of the prior parameters and $n = \sum_{i,j} n_{ij}$ is the total count of observed data [59]. For a single multinomial trial, the posterior predictive distribution is:

$$
\begin{aligned}
p(\hat{n}_{ij}|\mathbf{n}) &= \int p(\hat{n}_{ij}|\boldsymbol{\pi})\, p(\boldsymbol{\pi}|\mathbf{n})\, d\boldsymbol{\pi} \\
&= \int \pi_{ij}\, p(\pi_{ij}|\mathbf{n})\, d\pi_{ij} \\
&= E(\pi_{ij}|\mathbf{n}) \\
&= \frac{\alpha_{ij} + n_{ij}}{\sum_{i,j}(\alpha_{ij} + n_{ij})} \\
&= \frac{\alpha_{ij} + n_{ij}}{\alpha_0 + n}
\end{aligned}
$$

Another way to sample from the posterior predictive distribution involves considering each probability vector $\boldsymbol{\pi}^*$ from the posterior distribution and sampling a new dataset $\mathbf{n}^*$ from a multinomial likelihood function. The steps are:

1. Sample $\boldsymbol{\pi}^* \sim p(\boldsymbol{\pi}|\mathbf{n})$, where $\boldsymbol{\pi}^*$ are the probability vectors from the posterior distribution.

2. Sample $\mathbf{n}^* \sim \text{Multinomial}(n, \boldsymbol{\pi}^*)$.

3. The posterior predictive distribution is given by $(\mathbf{n}^*|\mathbf{n})$.

This posterior predictive sampling method originates from Bayesian statistics, where predictions about new data are made by integrating over the posterior distribution of the parameters [59]. The notation $p(\mathbf{n}^*|\mathbf{n})$ represents the posterior predictive distribution of new data $\mathbf{n}^*$ given the observed data $\mathbf{n}$.

5.3 Bayesian inference for test reproducibility

The reproducibility of a test is an important factor in determining the practical relevance of test results. Recently, there has been a lot of interest in the reproducibility probability (RP), which is not only estimated but also defined and interpreted differently in the classical frequentist statistics framework. Reproducibility provides an inference method for the probability for the event that, if repeated under identical circumstances and with the same sample size, the test outcome will be the same. Using Bayesian inference as detailed in Section 5.2, posterior predictive samples are drawn to evaluate reproducibility. Section 5.2 outlines the Bayesian framework for

Algorithm 5 Calculating Bayes-RP for tests of independence

```
1: Apply the tests of independence on the original sample, record the test outcome: $C^* = 1$ if $H_0$ is rejected and $C^* = 0$ if $H_0$ is not rejected.
2: Draw a posterior predictive sample based on the original sample and apply a test of independence.
3: Perform Step (2) $q$ times for $j = 1, \ldots, q$ and each time record the test decision: $C_j^* = 1$ if $H_0$ is rejected and $C_j^* = 0$ if $H_0$ is not rejected.
4: Calculate rp, where $rp = \sum_{j=1}^{q} \mathbb{1}(C^* = C_j^*)/q$.
```

a 2×2 contingency table, including the use of Dirichlet priors and multinomial likelihoods. In this section, we use the Bayesian inference to study the reproducibility of the chi-square test, likelihood ratio test, McNemar's test and Fisher's exact test, the same tests considered in Chapter 4 but from NPI perspective.

5.3.1 Bayesian inference for reproducibility of tests of independence

This section studies Bayesian inference for reproducibility of tests of independence. Algorithm 5 uses Bayesian inference to evaluate the reproducibility probability for test of independence, indicated by Bayes-RP. Goodman [40] employs a Bayesian approach using a non-informative prior to objectively assess statistical evidence without introducing subjective biases. In this study, we use a non-informative Dirichlet prior with parameters set to Dir(1,1,1,1). In Algorithm 5, the input is an original sample, and the number of posterior predictive samples $q = 1000$ per run.

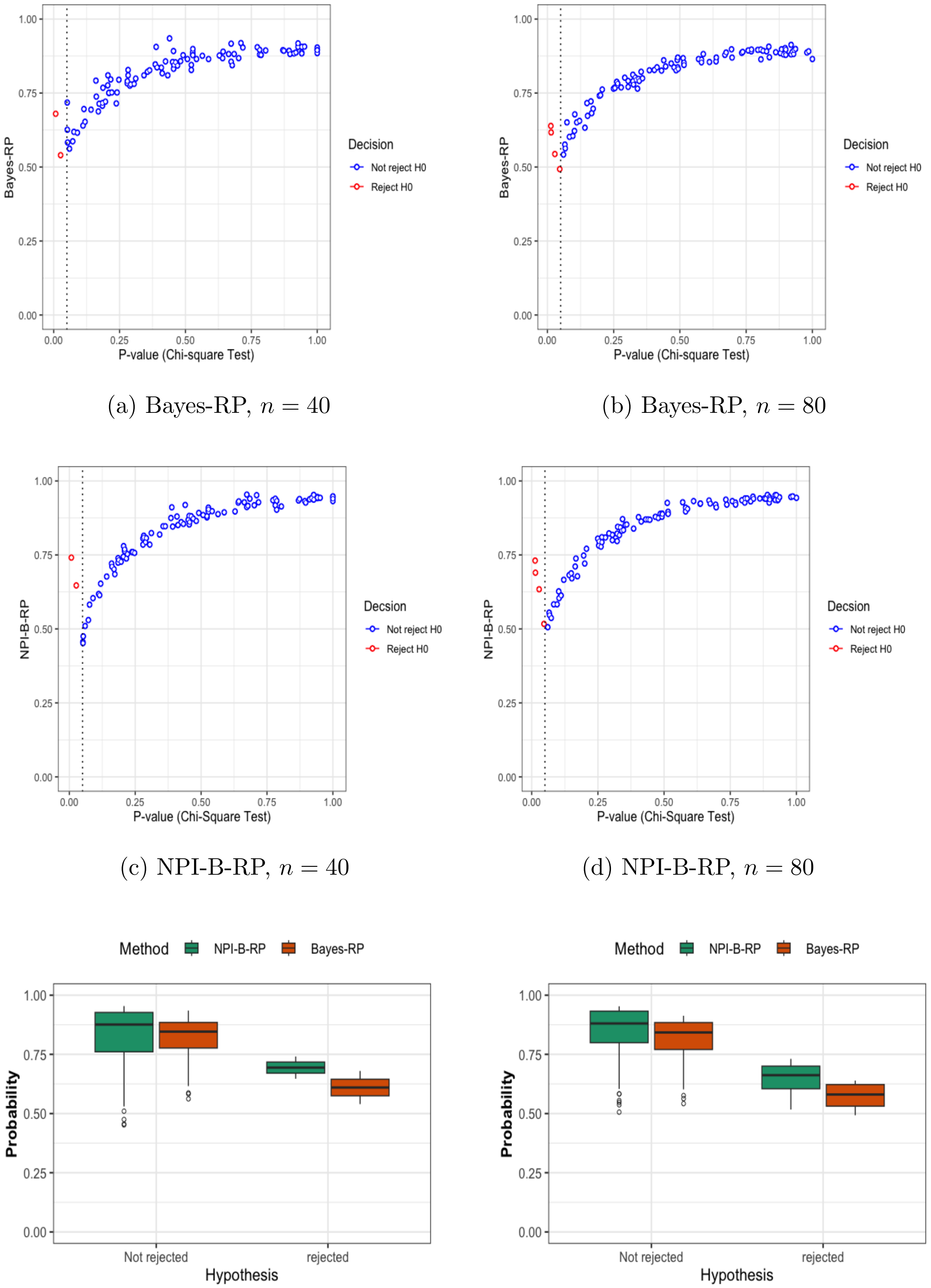
In this study, the reproducibility probability for the chi-square test and the likelihood ratio test of independence are investigated using the Bayesian approach and the NPI bootstrap approach. Data are generated using a multinomial distribution with probabilities $(0.18, 0.12, 0.42, 0.28)$ under H_0. The data are also generated using a multinomial distribution with two probability scenarios $(0.4, 0.2, 0.1, 0.3)$ and $(0.6, 0.1, 0.1, 0.2)$ under H_1. We conducted simulations under an alternative hypothesis to investigate whether two alternative distributions could lead to differing reproducibility results. In our case, we found no significant impact on reproducibility. The level of significance is 0.05. The simulation is considered with the sample sizes $n = 40, 80$. For a sample size of $n = 40$, the power of the test with a medium effect size $w = 0.3$ is approximately 0.47 , and with a large effect size $w = 0.5$, the power is approximately 0.88. For a sample size of $n = 80$, the power of the test with a medium effect size $w = 0.3$ is approximately 0.76 , and with a large effect size $w = 0.5$, the power is approximately 0.99. The Bayes-RP and the NPI-B-RP approach are applied using Algorithm 5 and Algorithm 4, respectively. The observed p-value and RP for both methods were determined for each of $N = 100$ samples. For the chi-square and likelihood ratio tests of independence, RP values are calculated using $B = 1000$ bootstrap samples and $q = 1000$ posterior predictive samples. Based on the bootstrap samples and the posterior predictive sample, RP values of the chi-square and likelihood ratio tests of independence are computed using the same original samples.

The relationship between reproducibility probability (RP) and the p-value for the chi-square and likelihood ratio tests is illustrated in Figures 5.1 and 5.2. These figures display RP values calculated using the NPI bootstrap and Bayesian methods under the null hypothesis (H_0) for sample sizes of 40 and 80. Boxplots of the RP values are presented for both NPI-B-RP and Bayes-RP, categorizing cases into rejection and non-rejection. In both methods, RP increases as the p-value moves away from the threshold of 0.05. Chapter 4 provides a detailed discussion on why reproducibility is lowest when the observed p-value is near this threshold. Simulation studies found that RP values based on NPI-B exhibit less variability compared to Bayesian methods, particularly with smaller sample sizes. This is because Bayesian methods are more influenced by the prior, leading to higher variability. Additionally, when sampling from the NPI-B, the categories are neighboring each other and have segment between each two categories. As the sample size n increases, the influence of the prior diminishes, resulting in similar RP outcomes between the Bayesian and NPI bootstrap methods for larger samples. When p-values range between 0.5 and 0.75, NPI-B-RP tends to be higher in non-rejection cases compared to Bayes-RP. Specifically, NPI-B-RP reaches values of 0.85 or higher within this p-value range, whereas Bayes-RP approaches approximately 0.85.

Under the alternative hypothesis H_1, two scenarios are investigated to examine the effect of various probabilities in cells on RP values for both methods. Figures 5.2, 5.3, 5.5, and 5.6 show RP values using bootstrap and Bayesian methods under H_1 for samples of sizes 40 and 80. The results show that as p-values get closer to 0.05, the behavior of RP values for NPI-B and Bayesian methods changes but stays similar between two scenarios. The power of the test is positively correlated with the sample size, which means that a larger sample size gives greater power. Increasing the sample size generally results in more cases where we reject the null hypothesis because the test becomes more powerful with larger samples. As shown in Figure 5.2, with $n = 40$, when the p-value is near 0.05, NPI-B-RP tends to be higher in rejection cases (shown in red) compared to non-rejection cases (blue). On the other hand, Bayes-RP is typically lower in rejection cases compared to nonrejection cases for p values close to 0.05. This behavior may be due to the influence of the prior in a Bayesian method and the sampling method from the NPI-B, where neighboring categories are adjacent and have segments between each pair of categories. As the sample size increases, we expect to observe more rejection cases, as illustrated in subfigures b and d of Figure 5.2. This is because the power of the test improves with larger sample sizes, leading to a higher probability of rejecting the null hypothesis when it is false. As the sample size increases to $n = 80$, both rejection and non-rejection cases with both methods approach 0.5 when p-values are close to the threshold of 0.05. Additionally, RP values become less variable. When the p-value is 0.01 and $n = 80$, NPI-B-RP is higher than Bayes-RP in rejection cases, with NPI-B-RP reaching around 0.75 and Bayes-RP around 0.65. However, the difference between the two remains small.

5.3.2 Bayesian inference for reproducibility of McNemar's test

In this section, the reproducibility of McNemar's test is studied using Bayesian and NPI-B methods. The null hypothesis is $H_0 : \pi_{12} = \pi_{21}$ and the alternative hypothesis is $H_1 : \pi_{12} \neq \pi_{21}$, the level of significance is 0.05. By following the same steps as for the tests of independence in Section 5.3.1, simulation studies are conducted to



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

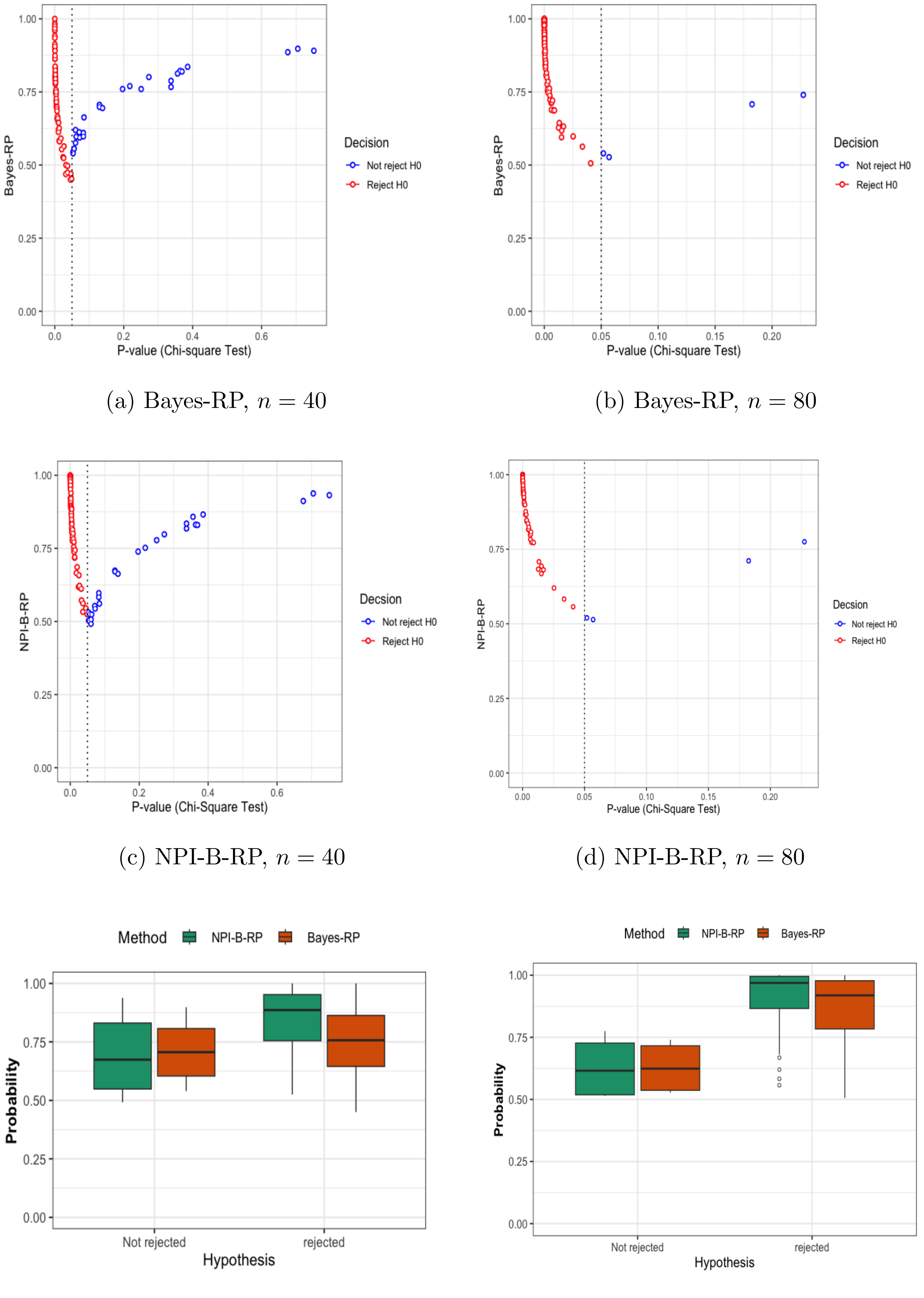
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.1: Simulations under H_0: Bayes-RP and NPI-B-RP values for chi-square test of independence.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

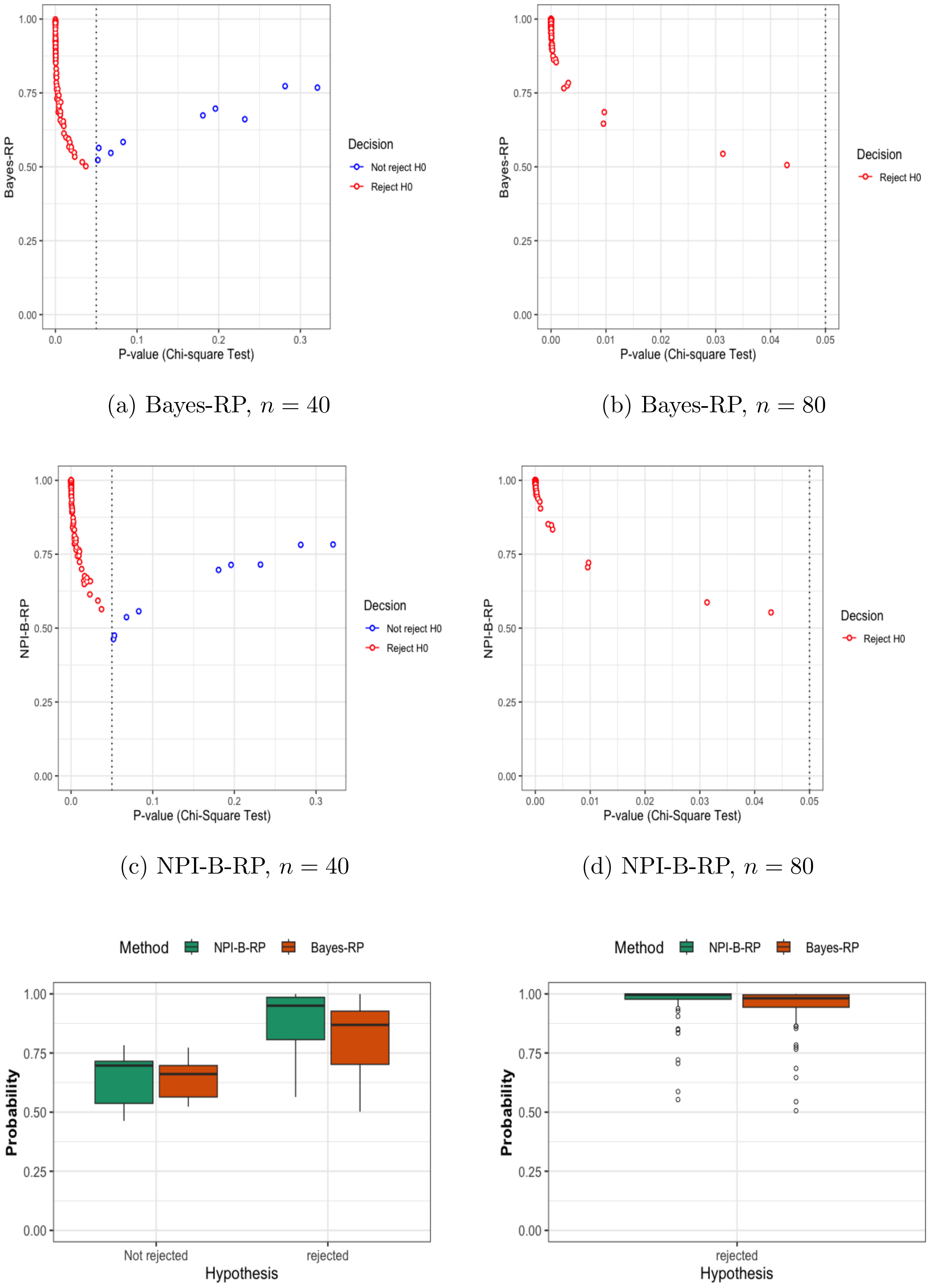
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.2: Simulations under H_1 with probabilities $(0.4, 0.2, 0.1, 0.3)$: Bayes-RP and NPI-B-RP values for chi-square test of independence.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

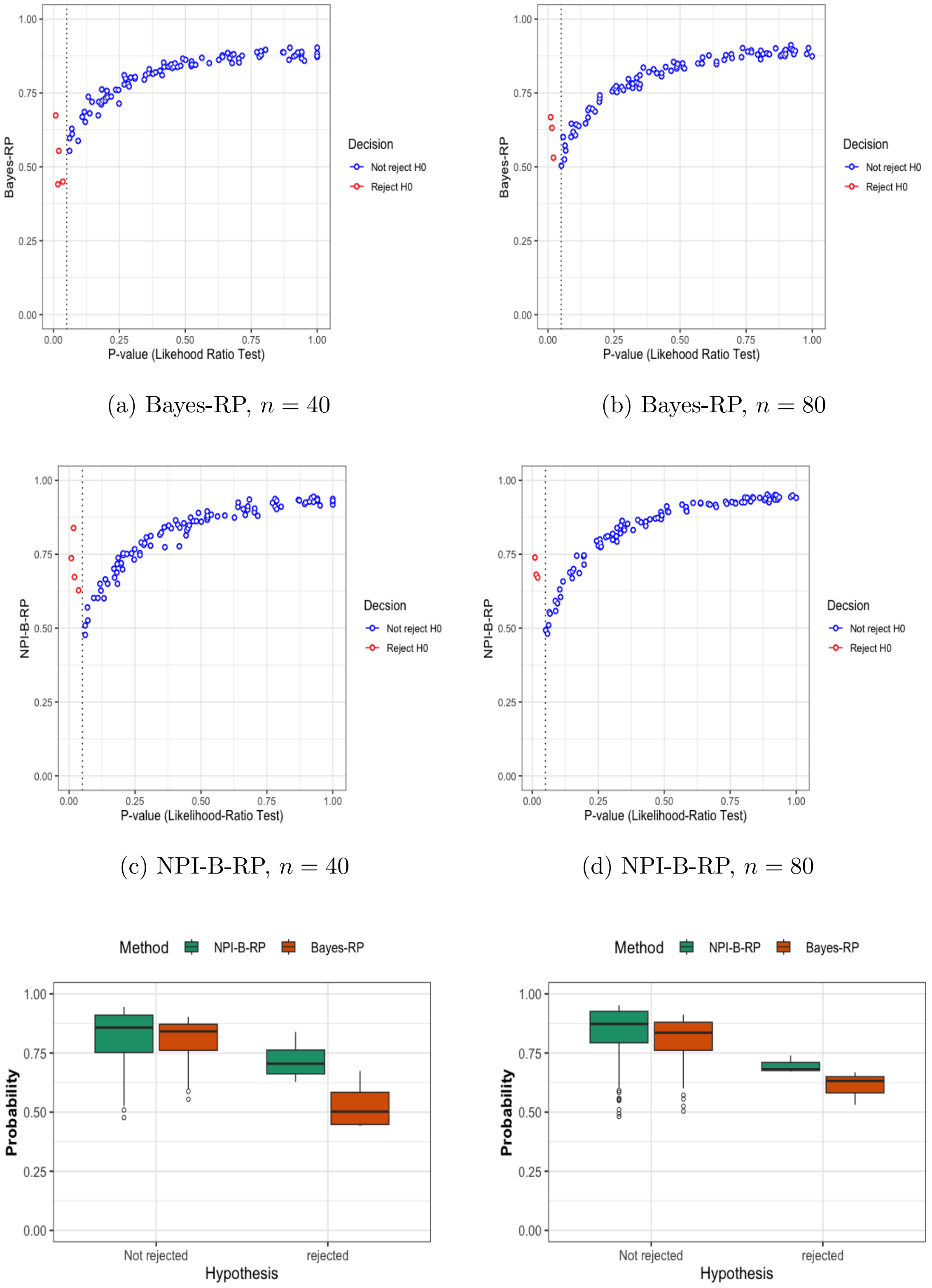
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.3: Simulations under H_1 with probabilities $(0.6, 0.1, 0.1, 0.2)$: Bayes-RP and NPI-B-RP values for chi-square test of independence.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

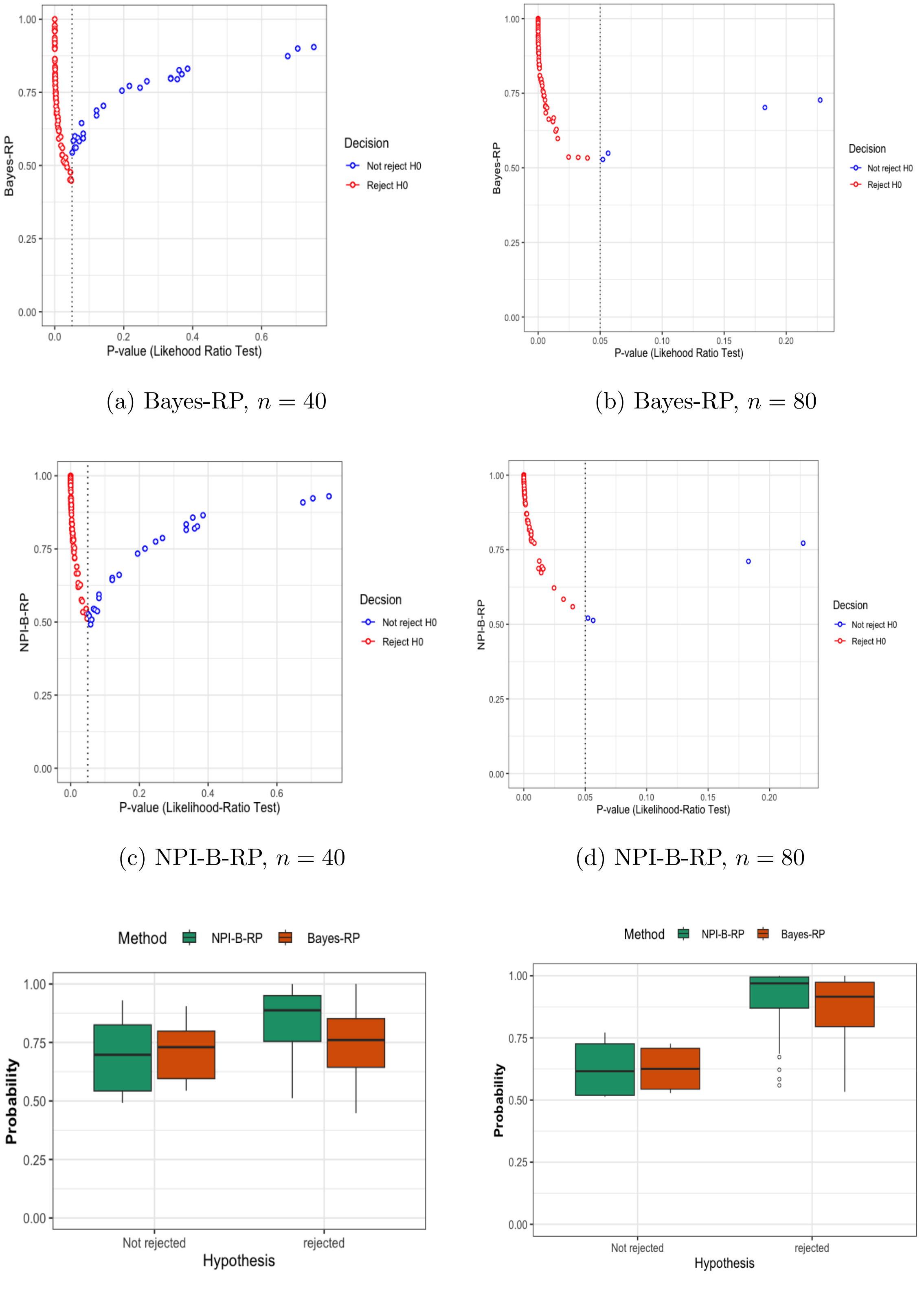
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.4: Simulations under H_0: Bayes-RP and NPI-B-RP values for likelihood ratio test of independence.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

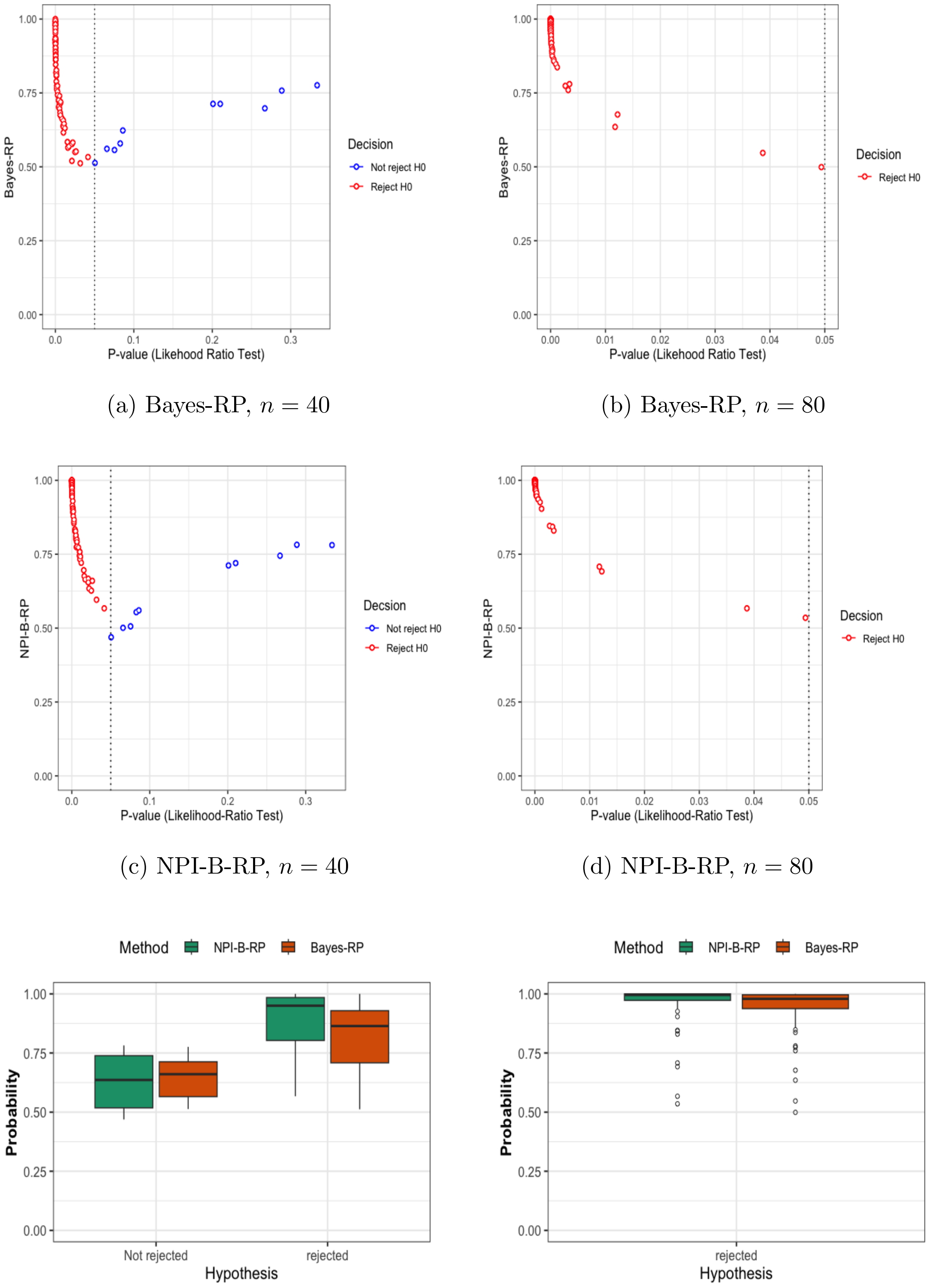
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.5: Simulations under H_1 with probabilities $(0.4, 0.2, 0.1, 0.3)$: Bayes-RP and NPI-B-RP values for likelihood ratio test of independence.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.6: Simulations under H_1 with probabilities $(0.6, 0.1, 0.1, 0.2)$: Bayes-RP and NPI-Bayes-RP values for likelihood ratio test of independence.

evaluate the Bayesian and NPI-B methods for RP of the McNemar test. The data are generated from the multinomial distributions with probabilities (0.2, 0.3, 0.3, 0.2) under H_0. The data are generated under H_1 from a multinomial distribution with probabilities (0.1, 0.2, 0.4, 0.3) and (0.1, 0.6, 0.1, 0.2). We simulate $N = 100$ samples of sizes $n = 40, 80$. McNemar's test RP value is computed using $B = 1000$ bootstrapping samples and $q = 1000$ posterior predictive samples.

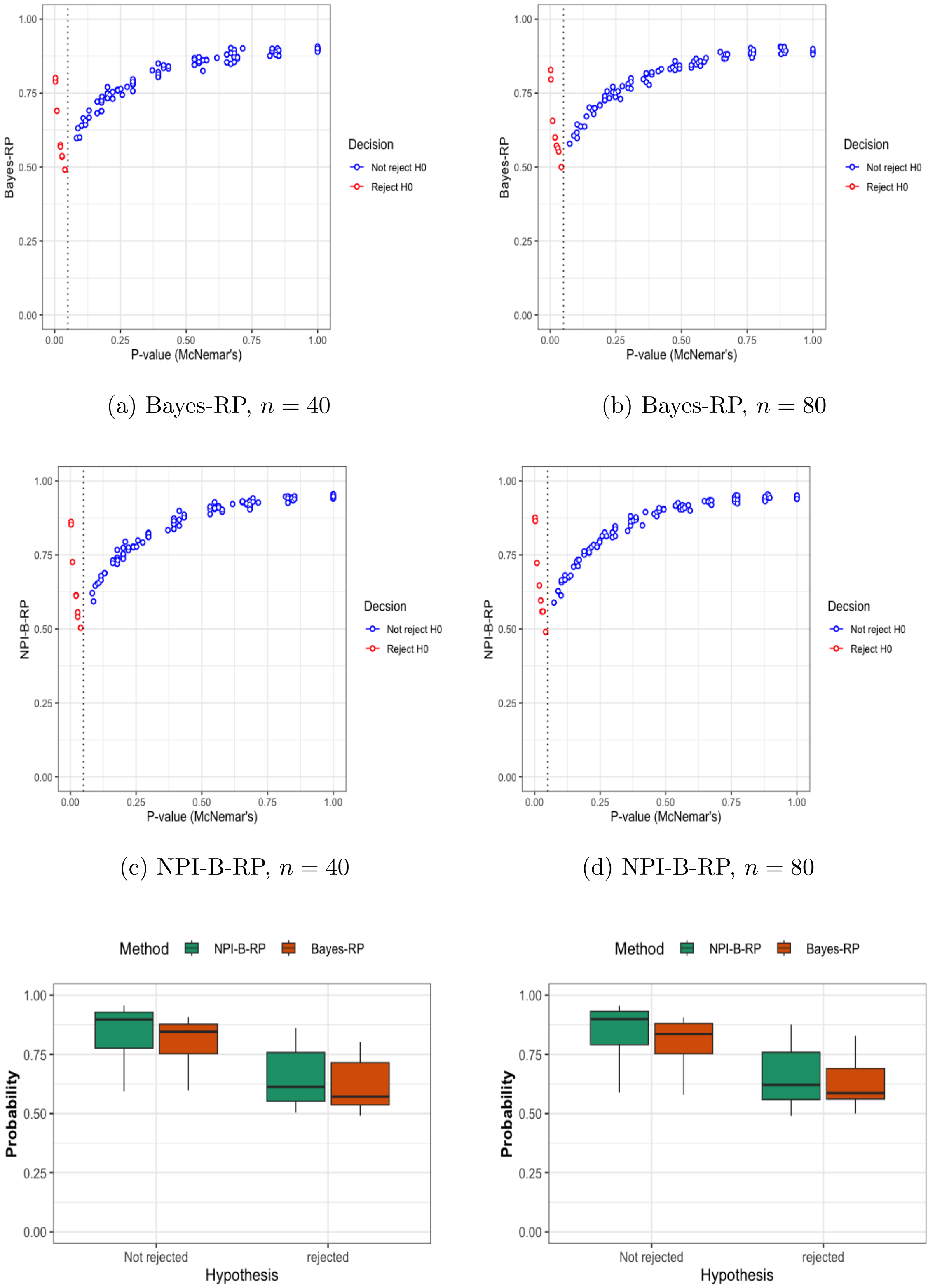
In the simulations, the RP for the McNemar's test is examined using NPI bootstrap and Bayesian methods. Figure 5.7 shows RP values using the bootstrap and Bayesian methods under H_0 for samples of size 40 and 80. The Bayes-RP and NPI-B-RP values tend to increase when the p-value is moved away from the threshold 0.05. As can be seen from the figures, reproducibility is at its lowest when the observed p-value is close to the threshold 0.05. Similar findings have been observed in previous studies of test reproducibility [11, 67], as well as in Chapter 4. There is a tendency for the Bayes-RP and NPI-B-RP to be lower in rejection cases (red cases) compared to non-rejection cases (blue cases) when the p-value is very close to the threshold 0.05. From a practical perspective, low values of RP are concerning, especially when H_0 is rejected with a p-value just below the level of significance, since many experiments explicitly aim to find evidence to support H_1. When sample sizes increase, both rejection and non-rejection RP values approach 0.5 when p-values are close to the threshold 0.05, as well as RP values become less variable. This occurs because increased sample size leads to an increase in the power of the test. The patterns of Bayes-RP and NPI-B-RP change when simulations are conducted under alternative hypotheses due to the change in p-value with respect to the threshold 0.05. When the p-value is 0.25, the NPI-B-RP pattern is higher in non-rejection cases compared to Bayes-RP. Specifically, NPI-B-RP is higher than 0.75 When the p-value is 0,25 , while Bayes-RP approaches 0.75.

Under the alternative hypothesis H_1, two scenarios are considered to study the impact of probabilities π_{12} and π_{21} on RP values for both methods. The power of test with cell probabilities $(0.1, 0.2, 0.4, 0.3)$ under the alternative hypothesis is approximately 0.372 for a sample size of $n = 40$ and approximately 0.637 for a sample size of $n = 80$. The power of test with cell probabilities (0.1, 0.6, 0.1, 0.2) under the alternative hypothesis is approximately 0.966 for a sample size of $n = 40$ and approximately 0.987 for a sample size of $n = 80$. Figures 5.8 and 5.9 show RP values using the bootstrap and Bayesian methods under H_1 for samples of size 40 and 80. The number of rejection cases increases when the difference between π_{12} and π_{21} increases, which is simply the result of the test becoming more powerful as the difference between π_{12} and π_{21} increases. Simulations under the alternative hypothesis result in a change in the pattern of RP values based on NPI-B and Bayesian methods, which reflect changes in the observed p-values with respect to the threshold 0.05. In both Bayesian and NPI bootstrap methods, similar results are observed for RP.

We compared the reproducibility of the chi-square test and McNemar's test using Bayesian and NPI-B methods. The data are generated from the multinomial distributions with probabilities (0.25, 0.25, 0.25, 0.25). The data are generated H_1 from a multinomial distribution with probabilities (0.1, 0.5, 0.2, 0.2). We simulate $N = 100$ samples of sizes $n = 40, 80$. The simulation results for the NPI-B-RP and Bayes-RP of the chi-square test and the McNemar's are provided in Appendix A. The McNemar's test produces similar results as the chi-square test of independence for reproducibility for both Bayesian and NPI-B methods. As n increases, the effect of the prior decreases, resulting in similar RP values for both methods. This convergence occurs because, with a larger sample size, the observed data has a greater influence on both methods, leading to similar outcomes.

5.3.3 Bayesian inference for reproducibility of Fisher's exact test

To evaluate reproducibility, Bayes-RP and NPI-B-RP for Fisher's exact test are studied using the same steps as for the tests of independence described in Section 5.3.1. The data are generated under H_0 from a multinomial distribution with probabilities (0.25, 0.25, 0.25, 0.25), whereas under the alternative hypothesis the data are generated from the multinomial distribution with probabilities (0.6, 0.1, 0.1, 0.2). For this simulation study, we use a small sample size of $n = 15, 40$ and $N = 50$ simulations per run. For each run, samples are generated from each of the multinomial



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

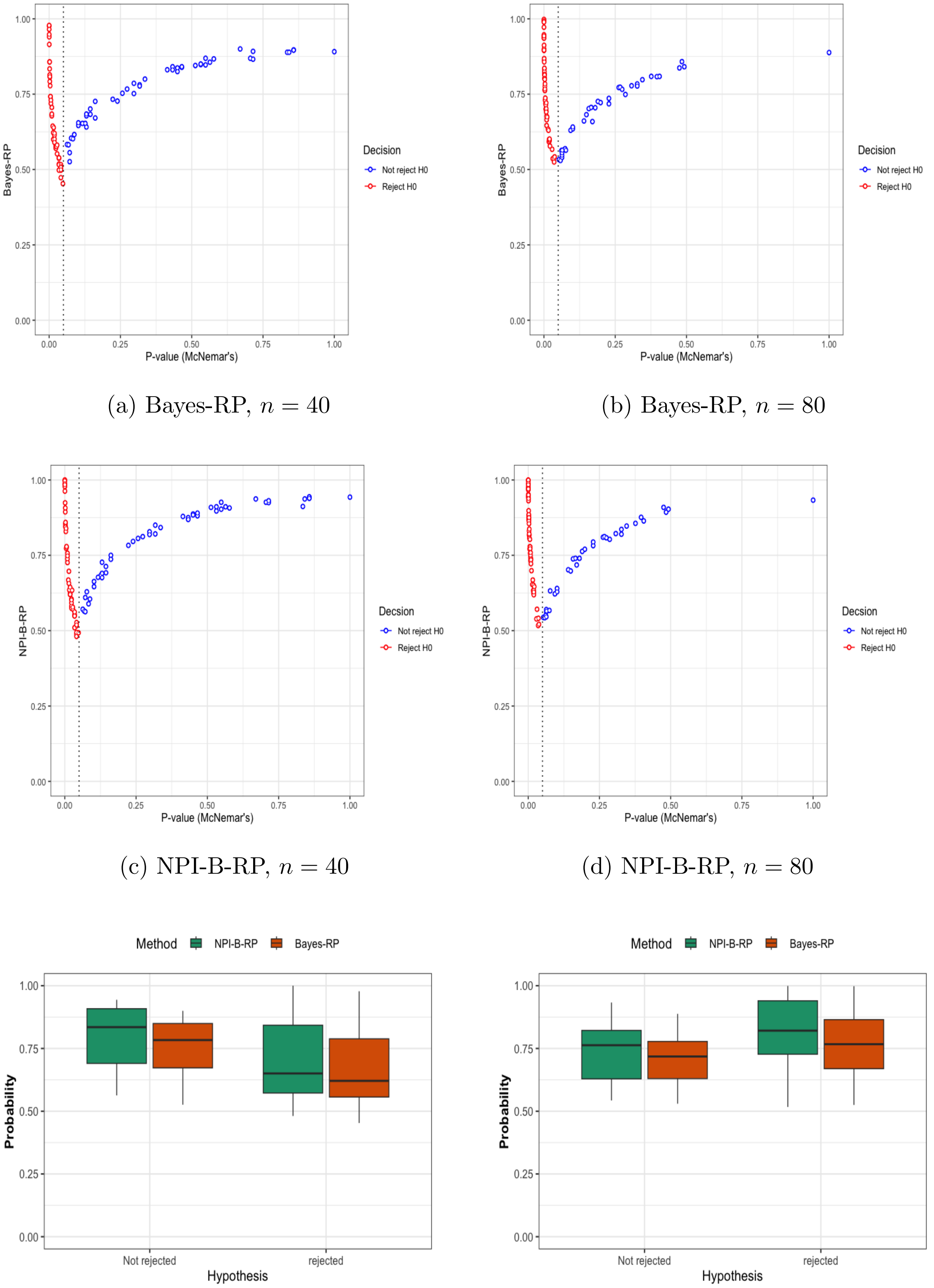
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.7: Simulations under H_0: Bayes-RP and NPI-B-RP values for McNemar's test.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

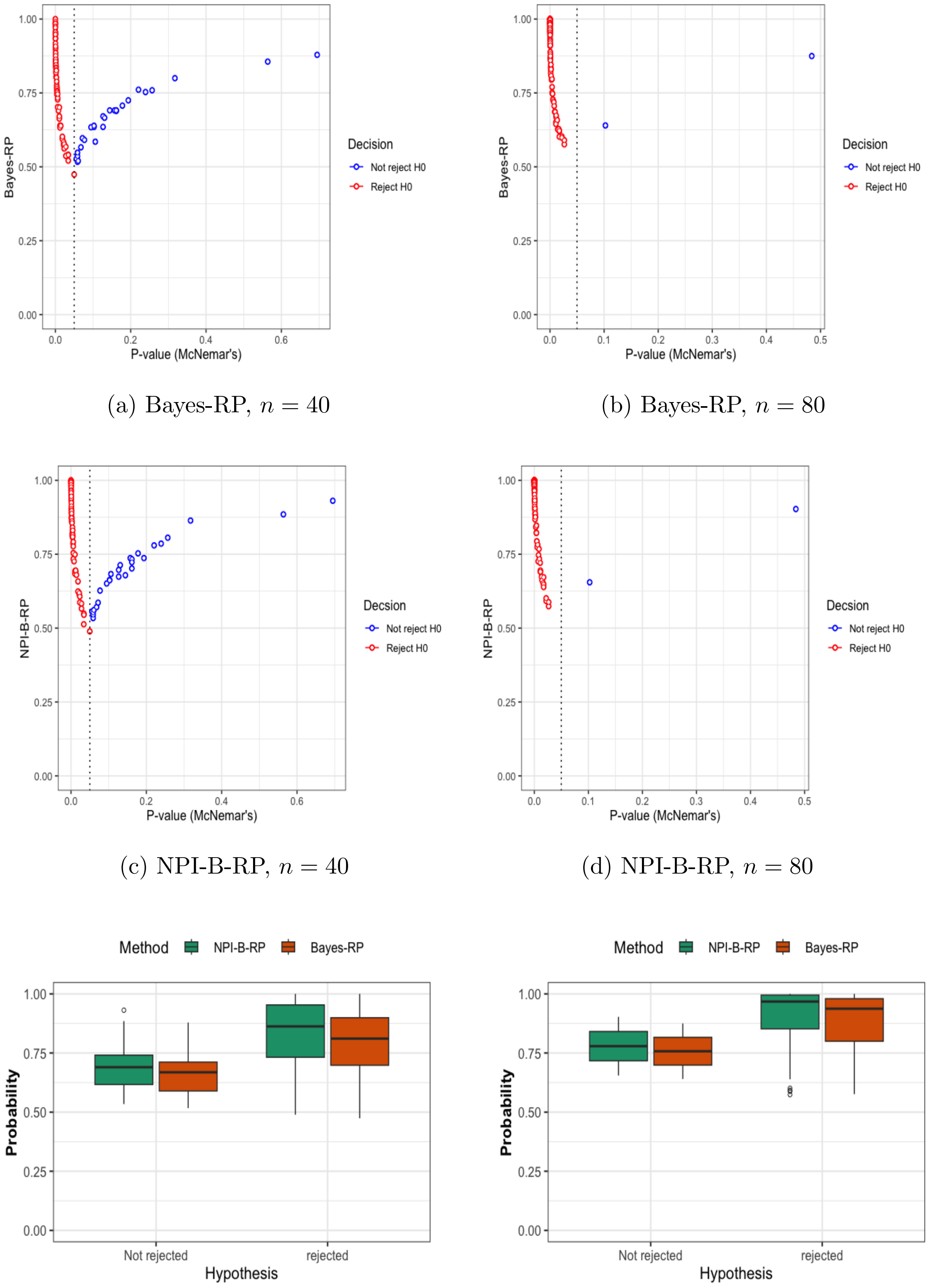
(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.8: Simulations under H_1 with probabilities $(0.1, 0.2, 0.4, 0.3)$: Bayes-RP and NPI-B-RP values for McNemar's test.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

(c) NPI-B-RP, $n = 40$

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

(e) RP, $n = 40$

(f) RP, $n = 80$

Figure 5.9: Simulations under H_1 with probabilities $(0.2, 0.6, 0.2, 0.2)$: Bayes-RP and NPI-B-RP values for McNemar's test.

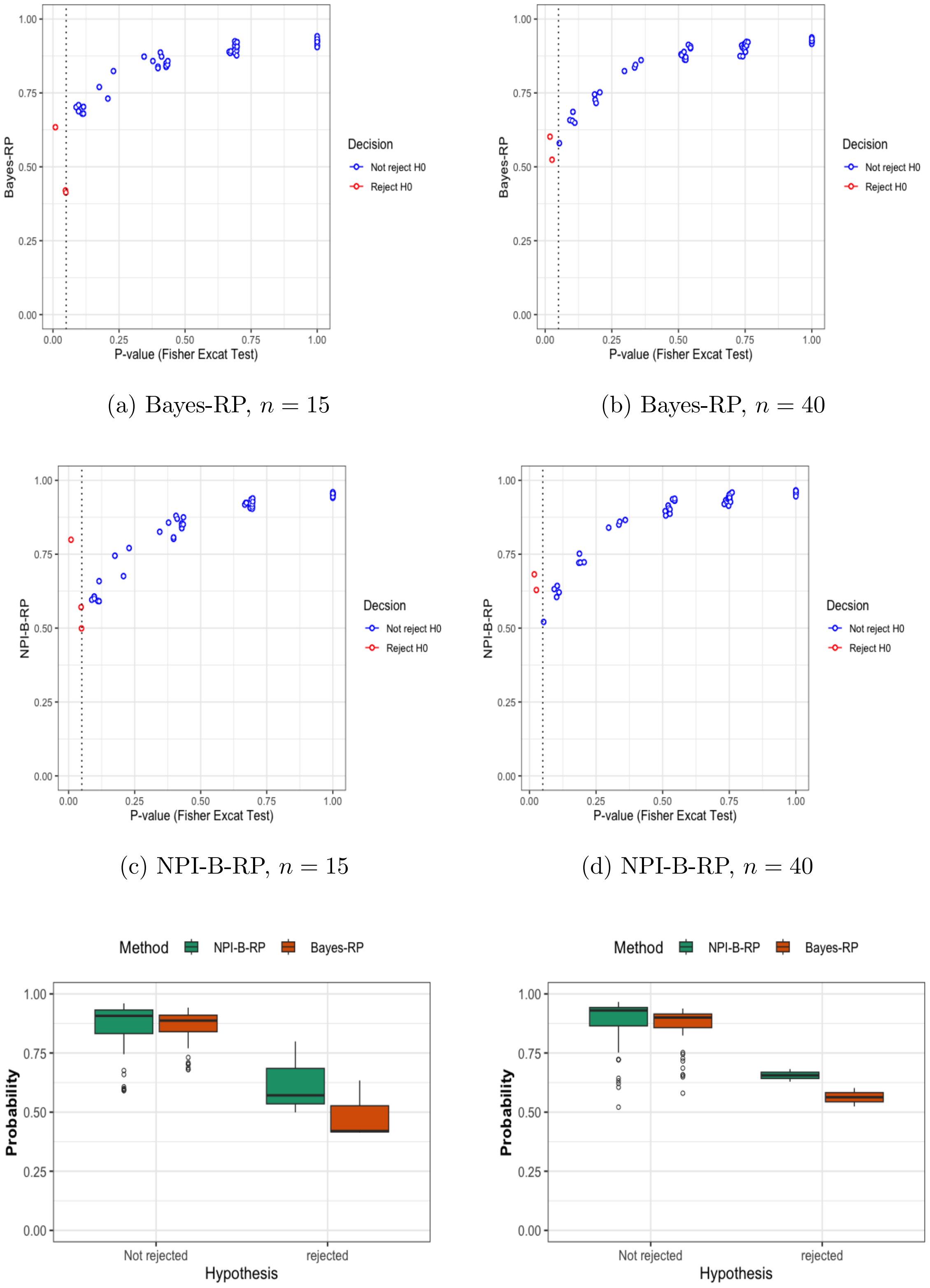
distributions, the Fisher's exact test of independence is conducted on the sample and the p-value is calculated. All values of RP were determined using $B = 1000$ bootstrap samples and $q = 1000$ posterior predictive samples.

The RP for the Fisher's exact test is examined based on Bayesian and NPI-B methods. Figures 5.10 and 5.11 display plots of these methods for simulations under H_0 and for simulations under H_1. Generally, the RP increases as the p-value moves away from the threshold 0.05. Reproducibility appears to be lowest around the threshold of the test as shown in the figures. Simulations indicate that RP values based on Bayesian and NPI-B have variability as a result of a small sample size. Simulations conducted under alternative hypotheses change the patterns of Bayes-RP and NPI-B-RP because the p-value changes in relation to the threshold 0.05. This occurs because increased sample size leads to an increase in the power test. In the simulations under H_1, increasing the sample size results in more rejection cases since a larger sample size makes the test more powerful.

A high NPI-B-RP value is observed in rejection cases (red cases in the figures) compared with non-rejection cases (blue cases) when the p-value is close to the threshold 0.05. Conversely, Bayes-RP value tends to be lower in cases of rejection than in non-rejection when the p-value is close to the threshold 0.05. When the sample size is small, the prior has a stronger effect on the results, causing increased variability. The RP for both methods tends to approach 0.5 when the observed p-value is close to the threshold 0.05 in both cases of rejection and non-rejection as the sample size increases. Additionally, an increase in sample size reduces the fluctuations in NPI-B-RP and Bayes-RP values.

5.4 Concluding remarks

This chapter presents the Bayesian method for evaluating reproducibility of statistical hypothesis tests based on 2×2 contingency tables. The Dirichlet prior was used in conjunction with a multinomial likelihood to drive the posterior predictive distribution, which is used to evaluate reproducibility of tests. In this chapter, the estimation of reproducibility is explored for tests of independence, Fisher's ex-



(a) Bayes-RP, $n = 15$

(b) Bayes-RP, $n = 40$

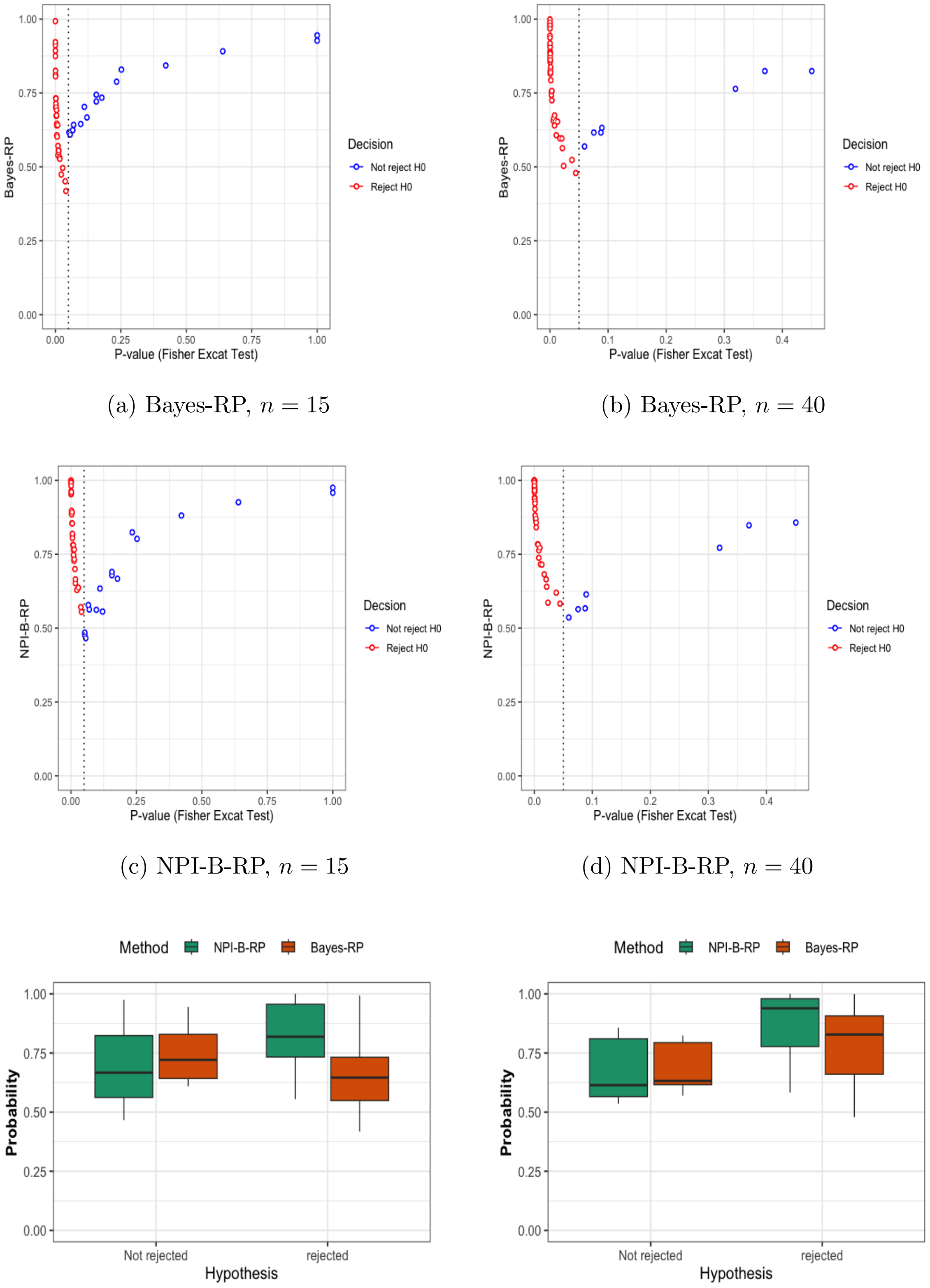
(c) NPI-B-RP, $n = 15$

(d) NPI-B-RP, $n = 40$

(e) RP, $n = 15$

(f) RP, $n = 40$

Figure 5.10: Simulations under H_0: Bayes-RP and NPI-B-RP values for Fisher's exact test.



(a) Bayes-RP, $n = 15$

(b) Bayes-RP, $n = 40$

(c) NPI-B-RP, $n = 15$

(d) NPI-B-RP, $n = 40$

(e) RP, $n = 15$

(f) RP, $n = 40$

Figure 5.11: Simulations under H_1: Bayes-RP and NPI-B-RP values for Fisher's exact test.

act test, and McNemar's test using Bayesian and NPI bootstrap methods. The Bayesian Inference and NPI-B are compared for the reproducibility of a variety of tests through the simulation studies. The test reproducibility is more naturally considered as a prediction problem than as an estimation problem. The NPI-B-RP is higher in rejection cases when the p-value is close to the threshold 0.05 than in non-rejection cases. In contrast, the Bayes-RP is generally lower in rejection cases than in non-rejection cases when the p-value is close to the threshold. As sample size increases, NPI-B-RP and Bayes-RP values become less fluctuation. A test's power generally increases with the size of the sample, so a larger sample size will increase power.

The reproducibility results for the tests of interest showed consistent results under various conditions, such as different sample sizes. However, other studies on test reproducibility outside of this thesis have reported different results, including cases where increasing the sample size does not affect the reproducibility patterns. To address this, we analyze each test using at least two sample sizes to investigate the impact of sample size on reproducibility. In our findings, we have not encountered unexpected reproducibility results for the tests of interest. RP tends to increase as the p-value increases.

As Senn [65] discussed, the circumstances in the real world may differ among different tests. It is possible to extend the bootstrap and Baysian methods for reproducibility of tests by using future sample sizes that differ from those in the data sample size or by using varying levels of statistical significance. From a theoretical perspective of reproducibility, it makes sense to utilize sample sizes and significance levels that are the same as those used in the actual test, particularly within a frequentist statistical framework.

Chapter 6

Reproducibility of Tests for Multiple 2×2 Tables

6.1 Introduction

NPI bootstrap and Bayesian methods for reproducibility probability for hypothesis tests based on the 2×2 tables are discussed in chapters 4 and 5. In this chapter, we explore the application of NPI bootstrap and Bayesian methods to assess the reproducibility of hypothesis tests based on multiple 2×2 tables. This chapter contributes to the development of NPI bootstrap and Bayesian methods for reproducibility by considering tests for the Mantel-Haenszel test, Breslow-Day test, and Woolf test.

The structure of the chapter is outlined as follows: Section 6.2 provides an overview of tests based on multiple 2×2 tables. Section 6.3 introduces the Bayesian and NPI bootstrap for reproducibility of the Cochran-Mantel-Haenszel test. Section 6.4 explores reproducibility for the Breslow-Day test using the NPI bootstrap and Bayesian methods. Section 6.5 introduces the Bayesian and NPI bootstrap for reproducibility of the Woolf test. The chapter ends with some concluding remarks in Section 6.6.

6.2 Tests based on multiple 2×2 tables

The presence of more than two cross-classification variables often results in multiple 2×2 tables in practice [48]. When exploring data in epidemiologic or clinical research, it is common to analyze multiple 2×2 contingency tables to assess the association between variables across different strata or subgroups [48]. For example, researchers may want to analyze the relationship between exposure (exposed/unexposed) and outcome (case/control) across different age groups, geographic regions, or other demographic factors.

In $2 \times 2 \times K$ tables, we have two binary characteristics X and Y, and an explanatory variable Z with K levels [48]. Table 6.1 displays the general notation for multiple 2×2 tables stratified by a third variable with K levels. For the partial frequency table n_{ijk} where $i, j = 1, 2$ and $k = 1, \ldots, K$, n_{i+k} represents the marginal total for row i at level k, n_{+jk} represents the marginal total for column j at level k, and n_k represents the total count of observations at level k. A partial table shows the cross-classification of two variables while keeping the third variable constant [48]. This means there are three possible sets of partial tables, depending on which variable is fixed at a specific level. The odds ratio for the corresponding marginal probabilities table can be defined as follows:

$$\theta_k^{XY} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}, \quad k = 1, \ldots, K$$

The estimated sample odds ratio is:

$$\hat{\theta}_k^{XY} = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}, \quad k = 1, \ldots, K$$

The variables X and Y might not be independent given Z, but the association between X and Y could be consistent across different levels of the conditioning variable. The null hypothesis is:

$$H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots = \theta_K^{XY} = \theta \qquad (6.2.1)$$

Under the null hypothesis H_0, all the odds ratios θ_k^{XY} across the K strata are equal to a common value θ. Conditional independence of X and Y given Z implies that there is no association between X and Y within each stratum of Z. Therefore,

	Y		
X	1	2	Total
1	n_{11k}	n_{12k}	n_{1+k}
2	n_{21k}	n_{22k}	n_{2+k}
Total	n_{+1k}	n_{+2k}	n_k

Table 6.1: Representation of multiple conditional 2×2 contingency tables for each level $K = k$ of the third variable Z.

setting $\theta = 1$ in the null hypothesis H_0 signifies that X and Y are conditionally independent given Z. Mathematically, if X and Y are conditionally independent given Z, then for each $k = 1, \ldots, K$:

$$\theta_k^{XY} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}} = 1$$

Thus, $H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots = \theta_K^{XY} = 1$ corresponds to the conditional independence of X and Y given Z. Mantel and Haenszel (1959) suggested an estimate for this common θ is [48] :

$$\hat{\theta}_{MH} = \frac{\sum_k \frac{n_{11k}n_{22k}}{n_k}}{\sum_k \frac{n_{11k}n_{21k}}{n_k}}$$

Various tests are commonly used to assess the relationship between variables across multiple 2×2 tables, including the Mantel-Haenszel test, the Breslow-Day test, and the Woolf test as described in Table 6.2. These methods allow researchers to determine if the association between two variables is consistent or differs significantly across levels in the data.

Table 6.2: Summary of Hypothesis Tests based on 2×2 Tables

Test	**Hypothesis**	**Assumptions**	**Properties**
Mantel-Haenszel Test	$H_0 : \theta_k^{XY} = 1,\ H_1 :$ $\theta_k^{XY} \neq 1,\ k = 1, \ldots, K$	Stratified into K subgroups; hypergeometric cell counts in each stratum.	Tests conditional independence; provides a common odds ratio across strata; chi-squared with 1 df.
Breslow-Day Test	$H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots =$ $\theta_K^{XY},\ H_1 :$ At least one θ_k^{XY} differs	Homogeneity of odds ratios under H_0; independent cell counts across strata.	Tests homogeneity of odds ratios across strata; chi-squared with $K - 1$ df.
Woolf Test	$H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots =$ $\theta_K^{XY},\ H_1 :$ At least one θ_k^{XY} differs	Logarithmic transformation stabilizes variance; accounts for small expected frequencies.	Uses weighted log odds ratios; chi-squared with $K - 1$ df.

6.2.1 Mantel-Haenszel Test

Cochran [17] investigated whether the success probability for two treatments for each of K contingency tables is the same. Mantel and Haenszel (MH) [55] introduced a test to examine the null hypothesis of conditional independence in $2 \times 2 \times K$ tables. The method they used was identical to Cochran's except for the correction factor associated with the finite population [76]. The Mantel-Haenszel statistic is an inference procedure to measure the association between two matched variables while controlling for a third variable. The MH test extends the usual χ^2 test while allowing stratification by a third variable.

Stratification refers to dividing the data into K homogeneous subgroups based on a third variable, to control for its potential confounding effect on the association between the two primary variables [2]. The MH test extends the usual chi-squared test by incorporating these strata, calculating the association within each stratum, and then combining the results to obtain an overall test statistic [55]. The hypotheses for the MH test are:

$$H_0 : \theta_k^{XY} = 1 \quad \text{vs} \quad H_1 : \theta_k^{XY} \neq 1, \quad k = 1, \ldots, K$$

In K stratified 2×2 tables, consider the row and column marginals of each of the K partial tables. In partial table k, the row totals are $\{n_{1+k}, n_{2+k}\}$, and the column totals are $\{n_{+1k}, n_{+2k}\}$. Given these totals, under H_0, the expected mean and variance of n_{11k} are derived from the hypergeometric distribution because, with fixed marginal totals, the cell counts follow a hypergeometric distribution in each stratum [17]. The hypergeometric mean and variance of n_{11k} are [2]:

$$\mu_{11k} = \frac{n_{1+k}\, n_{+1k}}{n_k}$$

$$\sigma^2_{11k} = \frac{n_{1+k}\, n_{2+k}\, n_{+1k}\, n_{+2k}}{n_k^2(n_k - 1)}$$

respectively. The cell counts from the different partial tables are independent of each other. We calculate the test statistic by adding up the differences between the observed and expected cell counts in all strata, adjusting each difference according to how much the data varies in that stratum (its variance) [55]. The Mantel-Haenszel test statistic is defined as:

$$MH = \frac{\left[\sum_k (n_{11k} - \mu_{11k})\right]^2}{\sum_k \sigma^2_{11k}}$$

The MH statistic follows a chi-squared distribution with 1 degree of freedom under H_0 because it tests a single parameter, the common odds ratio across all strata, against the null hypothesis that it equals one [2]. By allowing for stratification, the MH test provides a more accurate assessment of association by controlling for the stratifying variable, which might confound the relationship between the primary variables [76].

6.2.2 Breslow-Day Test

The Breslow-Day (BD) test is commonly employed to assess the homogeneity of odds ratios across multiple 2×2 contingency tables. This test examines whether the relationship between variables X and Y differs across various levels of the variable Z [14]. In other words, it tests whether the association between X and Y is consistent across different strata defined by Z, or if there is significant variation in the odds ratios between strata. The test statistic asymptotically follows a chi-squared distribution under the null hypothesis of homogeneity. The hypotheses for the Breslow-Day test are as follows:

$$H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots = \theta_K^{XY}$$

$$H_1 : \text{At least one of } \theta_k^{XY} \text{ is different from the others}$$

Here, θ_k^{XY} represents the odds ratio between X and Y in stratum k. Under the null hypothesis H_0, all strata share a common odds ratio, implying that the effect of X on Y is homogeneous across all levels of Z. Based on the conditional distribution of n_{11k} under H_0 given $\mathbf{n_c} = (n_{1+k}, n_{+1k}, n_k)$, which follows a noncentral hypergeometric distribution when $\theta \neq 1$, and considering the independence of n_{11k} across different strata [48]. Consider the mean $\mu_{11k}(\theta) = E(n_{11k}|\mathbf{n_c}, \theta)$ and variance $\sigma^2_{11k}(\theta) = \text{Var}(n_{11k}|\mathbf{n_c}, \theta)$ of n_{11k}. The $\mu_{11k}(\theta)$ is the estimated solution of the equation [48]:

$$\frac{\hat{\mu}_{11k}(n_{2+k} - n_{+1k} + \hat{\mu}_{11k})}{(n_{+1k} - \hat{\mu}_{11k})(n_{1+k} - \hat{\mu}_{11k})} = \theta \tag{6.2.2}$$

$$\hat{\sigma}^2_{11k}(\theta) = \left[\frac{1}{\hat{\mu}_{11k}} + \frac{1}{n_{2+k} - n_{+1k} + \hat{\mu}_{11k}} + \frac{1}{n_{+1k} - \hat{\mu}_{11k}} + \frac{1}{n_{1+k} - \hat{\mu}_{11k}}\right]^{-1} \tag{6.2.3}$$

The BD test statistic follows a χ^2 distribution with $K - 1$ degrees of freedom:

$$\chi^2_{BD} = \sum_{k=1}^{K} \frac{(n_{11k} - \hat{\mu}_{11k}(\hat{\theta}_{MH}))^2}{\hat{\sigma}^2_{11k}(\hat{\theta}_{MH})} \tag{6.2.4}$$

This statistic sums the squared differences between the observed and expected cell counts in each stratum, standardized by their variances. It measures the deviation of each stratum's odds ratio from the common odds ratio estimated by the Mantel-Haenszel estimator $\hat{\theta}_{MH}$. Here, $\hat{\mu}_{11k}(\hat{\theta}_{MH})$ and $\hat{\sigma}^2_{11k}(\hat{\theta}_{MH})$ are the expected value and variance evaluated by equations (6.2.2) and (6.2.3), respectively, replacing θ with the Mantel–Haenszel estimator $\hat{\theta}_{MH}$.

6.2.3 Woolf Test

The Woolf test assesses the homogeneity of odds ratios across multiple 2×2 contingency tables by applying a logarithmic transformation to the odds ratios before calculating the test statistic. Woolf [75] accounted for small expected cell frequencies to improve the validity of the usual χ^2 test. The Woolf test uses logarithms of the odds ratios to stabilize the variance and make the test statistic more normally distributed, which is helpful when sample sizes are small or events are rare [75]. To test the homogeneity of odds ratios using the Woolf test, consider the following hypotheses:

$$H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots = \theta_K^{XY}$$
$$H_1 : \text{At least one of } \theta_k^{XY} \text{ is different from the others}$$

In these hypotheses, θ_k^{XY} represents the odds ratio between X and Y in group k. Testing if they are all equal checks whether the effect of X on Y is the same in every group [48]. We calculate $\hat{\theta}_k$ and weights w_k as follows:

$$\hat{\theta}_k = \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}}, \quad k = 1, \ldots, K$$
$$w_k = \left(\frac{1}{n_{11k}} + \frac{1}{n_{12k}} + \frac{1}{n_{21k}} + \frac{1}{n_{22k}}\right)^{-1}, \quad k = 1, \ldots, K$$

The weights w_k are calculated so that groups with larger sample sizes and more precise estimates have more influence [48]. Next, we calculate the weighted average of $\log \hat{\theta}_k$ as follows:

$$\hat{\theta}_W = \exp\left(\frac{\sum_{k=1}^{K} w_k \log \hat{\theta}_k}{\sum_{k=1}^{K} w_k}\right)$$

The combined estimate $\hat{\theta}_W$ is a weighted average that represents the common odds ratio if all groups have the same effect [48]. The Woolf test statistic is defined as:

$$W = \sum_{k=1}^{K} w_k \left(\log \hat{\theta}_k - \log \hat{\theta}_W\right)^2 \tag{6.2.5}$$

Which follows the χ^2 distribution with $K - 1$ degrees of freedom [48]. If the Woolf test statistic is large, it means the odds ratios are different across groups, indicating that the effect of X on Y changes with the stratifying variable Z [48].

Algorithm 6 Reproducibility for the MH test

```
1: Apply the MH test on $K$ table, and make decision about $H_0$, then record the test outcome: $C^* = 1$ if $H_0$ is rejected and $C^* = 0$ if $H_0$ is not rejected.
2: Draw NPI-B sample and posterior predictive sample from each $K$ table based on the original sample of $K$ table and apply a MH test and obtain the results.
3: Repeat Step 2 $B$ times for $j = 1, \dots, B$ and $q$ times for $j = 1, \dots, q$, and each time record the test decision: $C_j^* = 1$ if $H_0$ is rejected, and $C_j^* = 0$ if $H_0$ is not rejected.
4: Calculate the rp, where $rp = \sum_{j=1}^{q} \mathbb{1}(C^* = C_j^*)/q$ and $rp = \sum_{j=1}^{B} \mathbb{1}(C^* = C_j^*)/B$.
```

6.3 Reproducibility for Mantel-Haenszel test

In this section, we compare two approaches: NPI-B and Bayesian methods for investigating the relative performance (RP) of Mantel-Haenszel test using Algorithm 6. The reproducibility of tests is naturally viewed as a prediction problem instead of an estimation problem, which is aligned well with these approaches. The explicit predictive properties of both NPI-B and Bayesian inference methods provide suitable frameworks for deducing the reproducibility of the MH test. To enable this comparison, we conduct simulation studies to assess the performance of NPI-B and Bayesian in evaluating the RP of the MH test under different scenarios.

The algorithm for reproducibility of the MH test begins by applying the MH test on K tables and making a decision about the null hypothesis H_0. The outcome is recorded as $C^* = 1$ if H_0 is rejected, or $C^* = 0$ if H_0 is not rejected. Next, an NPI-B sample and a posterior predictive sample are drawn from each K table independently and the MH test is applied again to obtain the results. This process is repeated B times for each iteration $j = 1, \dots, B$, and q times for each iteration $j = 1, \dots, q$, with the decision recorded each time. The recorded decisions are $C_j^* = 1$ if H_0 is rejected, and $C_j^* = 0$ if H_0 is not rejected. Finally, the reproducibility is calculated by computing the ratio of matching decisions across all test outcomes. This is expressed as $rp = \sum_{j=1}^{q} \mathbb{1}(C^* = C_j^*)/q$, and $rp = \sum_{j=1}^{B} \mathbb{1}(C^* = C_j^*)/B$. The algorithm is as follows:

The purpose of this study is to examine the RP for the MH test using the NPI-B and the Bayesian methods. RP values are calculated using bootstrap samples $B = 1000$ and posterior predictive samples $q = 1000$. The same prior in chapter 5 was used. This study will focus on $K = 3$, where K represents the number of tables. Simulations were conducted under both the null hypothesis H_0 and the alternative hypothesis H_1, with $N = 50$ runs per simulation. Data were generated using a multinomial distribution based on specified odds ratios. Under H_0, the odds ratios θ_k^{XY} were set to 1 for all k. Under H_1, the odds ratios were set to different values, with $\theta_1^{XY} = 2$, $\theta_2^{XY} = 2$, and $\theta_3^{XY} = 3$. In all figures in this chapter, red circles represent rejection cases of H_0, while blue circles represent non-rejection cases.
In $K = 3$, we consider three different scenarios depending on the sample size: the first scenario considers small size in each K tables, the second Scenario considers large size in each K tables, and the third scenario considers small or large size in each K tables. Three scenarios with different sample sizes. In scenario 1, with sample sizes $n_1 = 60$, $n_2 = 50$, and $n_3 = 40$, Figure 6.1 displays the RP values using NPI bootstrap and Bayesian methods under both H_0 and H_1. Boxplots illustrate the cases of rejection and non-rejection for both methods. Scenario 2, with sample sizes $n_1 = 160$, $n_2 = 140$, and $n_3 = 120$, is depicted in Figure 6.2, showing the RP values and boxplots for both RP methods under H_0 and H_1. Finally, Scenario 3, with sample sizes $n_1 = 160$, $n_2 = 120$, and $n_3 = 60$, is illustrated in Figure 6.3, presenting the RP values and boxplots for both RP methods under H_0 and H_1. In all scenarios, the RP values are obtained using NPI bootstrap and Bayesian methods under the null and alternative hypotheses. The boxplots compare the RP values based on the NPI-B-RP and Bayes-RP methods when the null hypothesis is rejected versus not rejected.

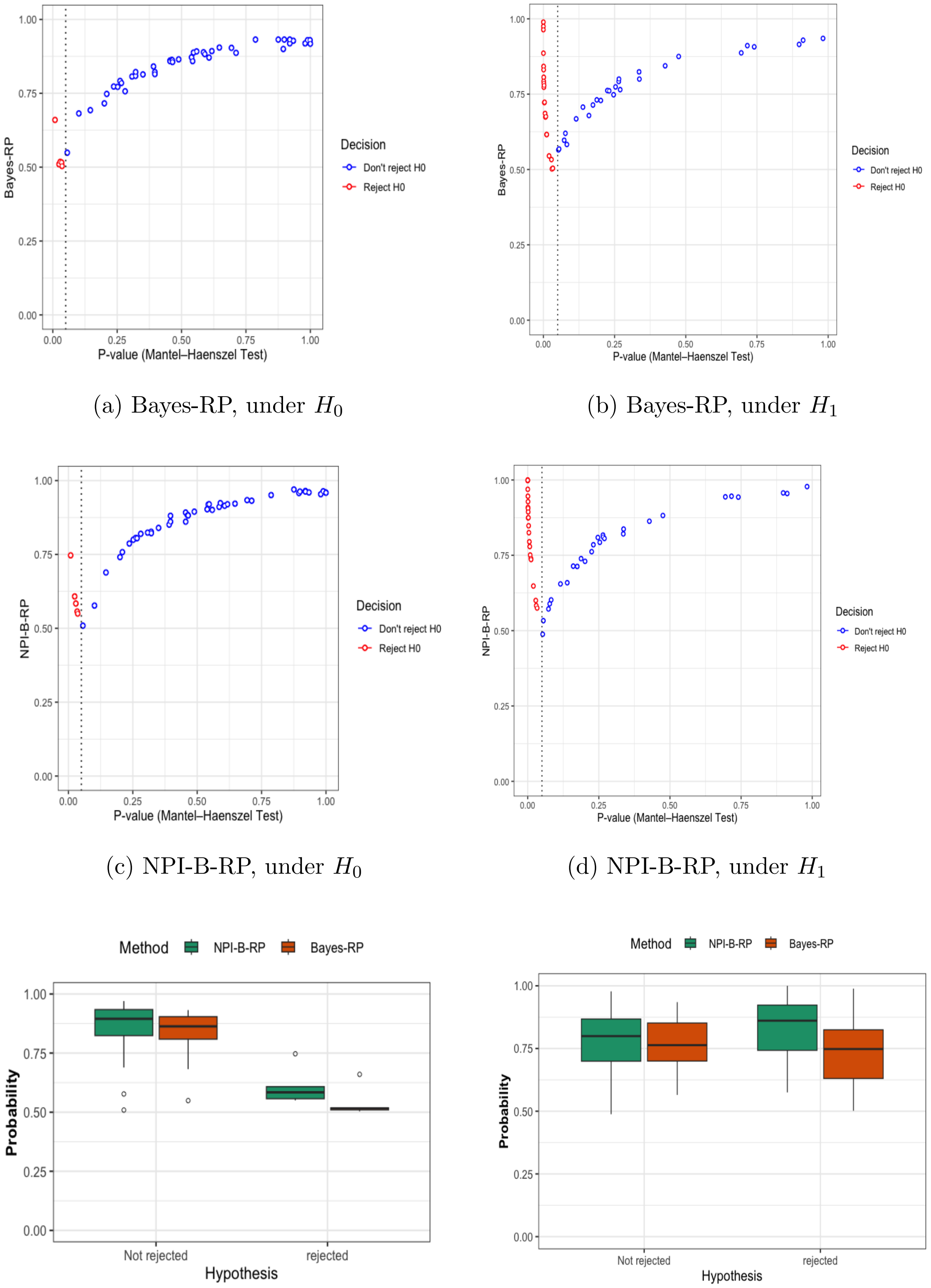
The difference appears clearly in scenario 2, since we consider that each K table is large, which is reflected in the low fluctuation in RP values in both methods. For both methods, the scenario 1 has the greatest variability in RP values, followed by scenario 3. According to the two methods, RP values tend to increase when the test statistic moves away from the test threshold. As expected, reproducibility is low near the test threshold, so if the p-value is close to 0.05, reproducibility is low. In these cases, it is uncertain whether the null hypothesis should be rejected because the evidence is not strong. As sample size increases, both rejection and non-rejection probabilities tend to converge towards 0.5 when p-values are close to the threshold 0.05, while the variability of RP values diminishes, mainly due to the increased the power test associated with larger sample sizes. By sampling under the alternative hypothesis, larger samples increase the power of the test, so more cases reject H_0.

When the p-value is 0.5 in all scenarios, the NPI-B-RP pattern is higher in non-rejection cases than Bayes-RP. Specifically, NPI-B-RP consistently achieves values greater than or equal to 0.85 within this p-value range, while the Bayes-RP approaches do not consistently reach 0.85. This difference arises because the prior information influences the Bayesian results more significantly than the data itself, resulting in a less consistent performance in non-rejection cases for Bayes-RP. The same pattern is observed in rejection cases, where NPI-B-RP also performs better than Bayes-RP. The results observed with $K = 5$ are similar to those observed with $K = 3$ and are provided in Appendix A.

This study investigated the impact of increasing the sample size on the behavior of reproducibility. The reproducibility results for the tests of interest showed consistent results under various conditions, such as different sample sizes. However, other studies on test reproducibility outside our work have reported different results, including cases where increasing the sample size does not affect the reproducibility patterns.

6.4 Reproducibility for Breslow-Day test

This section examines the RP of the Breslow-Day test using NPI-B and Bayesian methods. Simulation studies are conducted to evaluate the Bayesian and NPI-B methods for RP of the Breslow-Day test using the same steps as for the MH test described in Section 6.3. The RP values are calculated by using bootstrap samples $B = 1000$ and posterior predictive samples $q = 1000$. In this study, we will focus on two cases where $K = 3$ and $K = 5$ to examine the impact of increasing the number of K tables on the RP. We run simulations under both H_0 and H_1, with $N = 50$ runs per



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

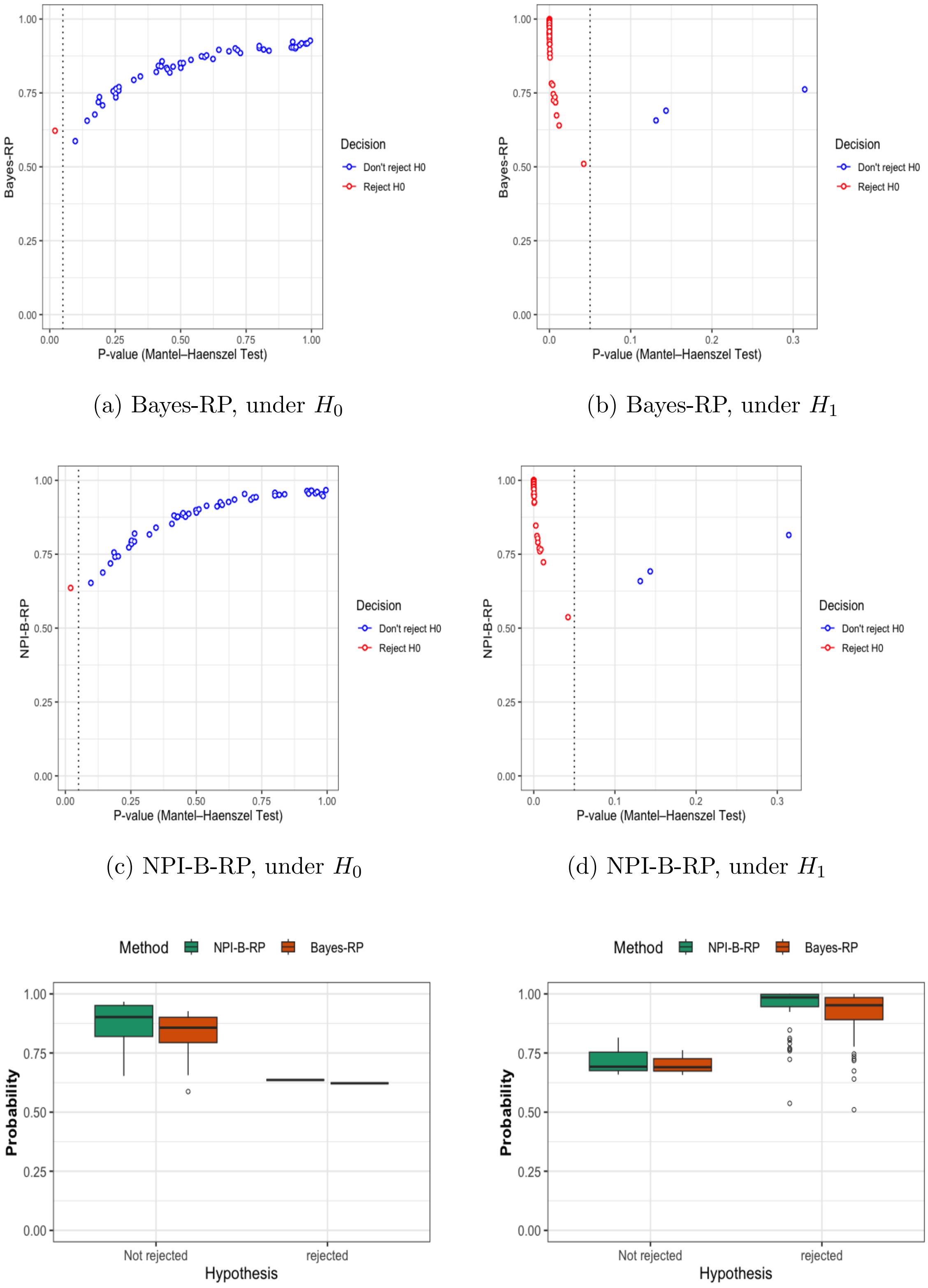
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.1: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Mantel-Haenszel test, Scenario 1 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

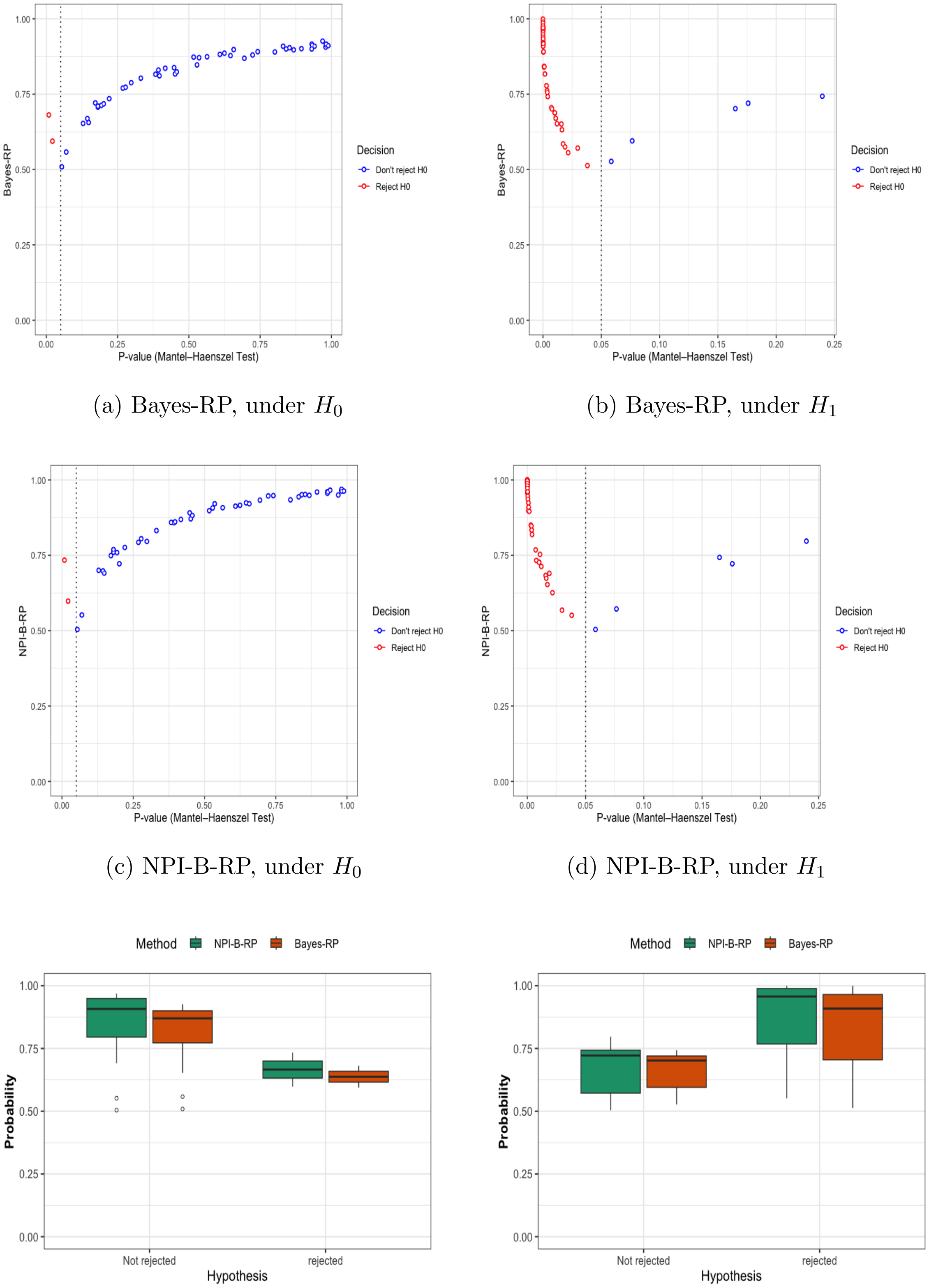
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.2: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Mantel-Haenszel test, Scenario 2 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.3: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Mantel-Haenszel test, Scenario 3 for $K = 3$.

simulation. The data are generated using odds ratios $H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots = \theta_K^{XY}$ under the null hypothesis, and H_1 : At least one of θ_k^{XY} is different from the others under the alternative hypothesis.

We consider three different scenarios for the case with $K = 3$, based on sample sizes: small sizes, large sizes, and a mix of small and large sizes. Data were generated using a multinomial distribution based on specified odds ratios: under H_0, $\theta_1^{XY} = \theta_2^{XY} = \theta_3^{XY} = 1$, and under H_1, at least one of θ_k^{XY} is different from the others, with $\theta_1^{XY} = 6$, $\theta_2^{XY} = 1$, and $\theta_3^{XY} = 2$. In scenario 1, the sample sizes are $n_1 = 60$, $n_2 = 50$, and $n_3 = 40$. Figure 6.4 shows the RP values using NPI-B-RP and Bayes-RP methods with bootstrapping and posterior predictive samples under both hypotheses, including boxplots for rejections and non-rejections. In scenario 2, with sample sizes $n_1 = 150$, $n_2 = 180$, and $n_3 = 160$, Figure 6.5 displays the RP values and boxplots for both methods. For scenario 3, where the sample sizes are $n_1 = 150$, $n_2 = 120$, and $n_3 = 50$, Figure 6.6 presents the RP values and boxplots for both hypotheses, showing both rejection and non-rejection scenarios. At a significance level of 0.05 and with 2 degrees of freedom, a critical value of 5.99 is calculated. In both methods, RP tends to increase when the test statistic moves away from the threshold 0.05. The reproducibility is the lowest close to the the threshold 0.05, as expected. By increasing the sample size, the variability of both methods decreases and the power of the test increases. In both methods, the RP tends to be higher in cases of rejection (red cases in the figures) than in cases of non-rejection (blue cases) when the p-value is close to the threshold 0.05. As a result of sampling under the alternative hypothesis, larger samples are more likely to reject H_0, due to increasing the power of the test. When the p-value is 0.2 in all scenarios, NPI-B-RP shows a higher in non-rejection cases compared to Bayes-RP. Specifically, NPI-B-RP reaches values of 0.65 or higher, while Bayes-RP approach 0.65.

In the second case with $K = 5$, three scenarios are identified based on the data generated. At a 0.05 significance level and 4 degrees of freedom, the critical value is 9.488. Under H_0, the odds ratios were $\theta_1^{XY} = \theta_2^{XY} = \theta_3^{XY} = \theta_4^{XY} = \theta_5^{XY} = 1$. Under H_1, at least one of the odds ratios differs, specifically $\theta_1^{XY} = 6$, $\theta_2^{XY} = 0.67$, $\theta_3^{XY} = 2$, $\theta_4^{XY} = 1$, and $\theta_5^{XY} = 1$. Scenario 1 has sample sizes $n_1 = 40$, $n_2 = 50$,

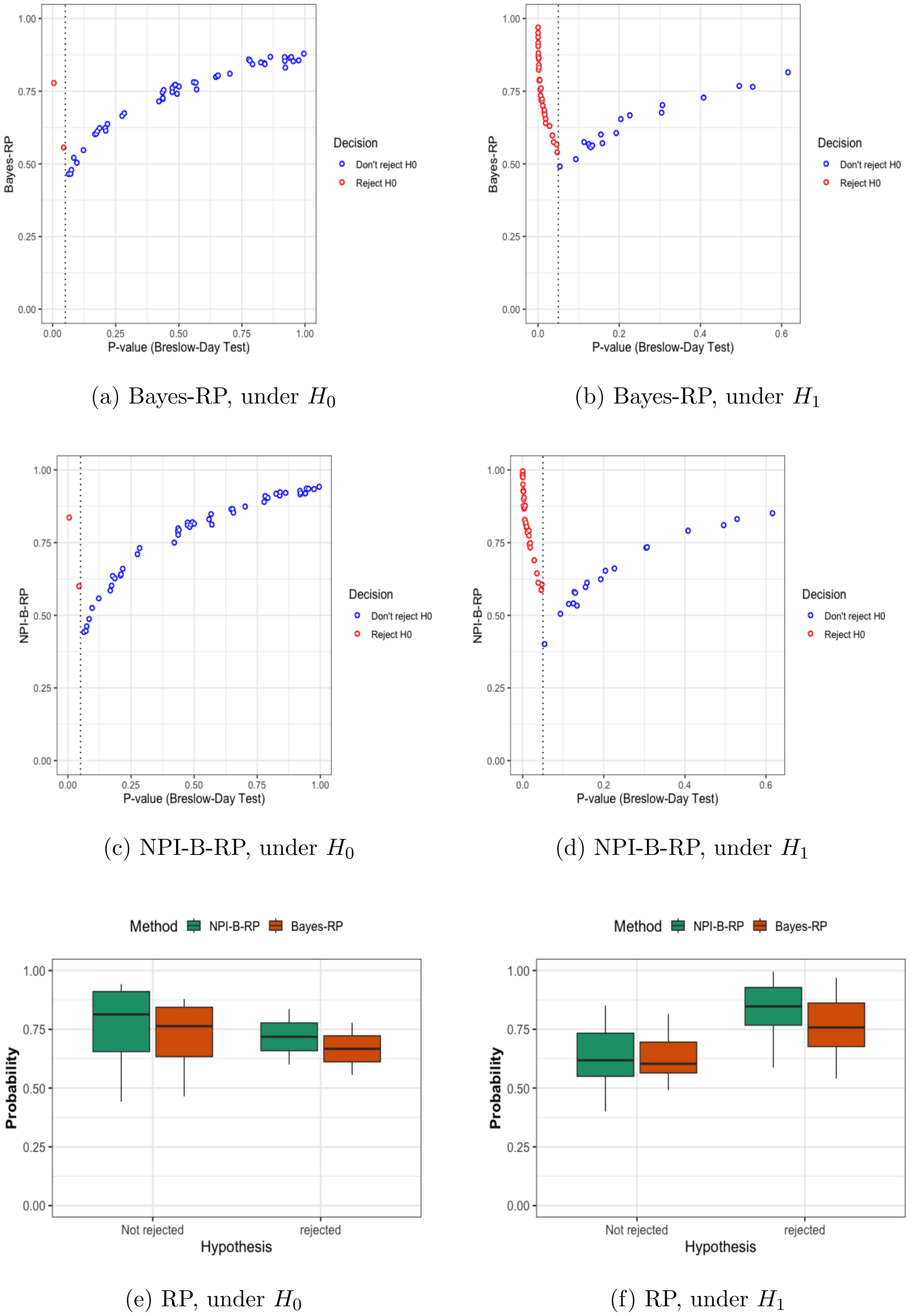
$n_3 = 55$, $n_4 = 60$, and $n_5 = 70$. Figure 6.7 shows the RP values using NPI-B-RP and Bayes-RP methods with bootstrapping and posterior predictive samples, including boxplots for both rejections and non-rejections. Scenario 2, with sample sizes $n_1 = 180$, $n_2 = 170$, $n_3 = 165$, $n_4 = 160$, and $n_5 = 150$, is illustrated in Figure 6.8, displaying RP values and boxplots for both hypotheses. In scenario 3, the sample sizes are $n_1 = 160$, $n_2 = 150$, $n_3 = 130$, $n_4 = 60$, and $n_5 = 40$. Figure 6.9 presents the RP values and boxplots under both hypotheses, showing rejection and non-rejection cases. When the p-value is 0.75 with $K = 5$ in all scenarios, NPI-B-RP shows a higher in non-rejection cases compared to Bayes-RP. Specifically, NPI-B-RP reaches values of 0.75 or higher, while Bayes-RP approaches 0.75. In comparison, with $K = 3$ in all scenarios with the same p-value, NPI-B-RP consistently achieves values greater than or equal to 0.85. At the same time, Bayes-RP approaches does not consistently reach 0.85.

The patterns of RP values appear to be affected by the increase in the number of K tables. As the number of K tables increases, both methods tend to show higher RP values in rejection cases (red cases) compared to non-rejection cases (blue cases) when the test statistic is close to the threshold.

The test statistic for the Breslow-Day test is computed using Equation (6.2.4). In multiple tables, the Breslow-Day test examines the homogeneity of odds ratios across all K tables. The overall test statistic is the sum of the individual Breslow-Day test statistics from each of the K tables. As the number of tables increases, the combined test statistic is more likely to exceed the critical value, thereby increasing the chances of rejecting the null hypothesis H_0. Consequently, the RP values of both methods tend to be lower in non-rejection cases and higher in rejection cases.

6.5 Reproducibility for Woolf test

The purpose of this section is to examine the RP of the Woolf test using NPI-B and Bayesian methods. By following the same steps as for the Mantel-Haenszel test in Section 6.3, simulation studies are conducted to evaluate the Bayesian and NPI-B methods for RP of the Woolf test. We calculate the RP values using bootstrap



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

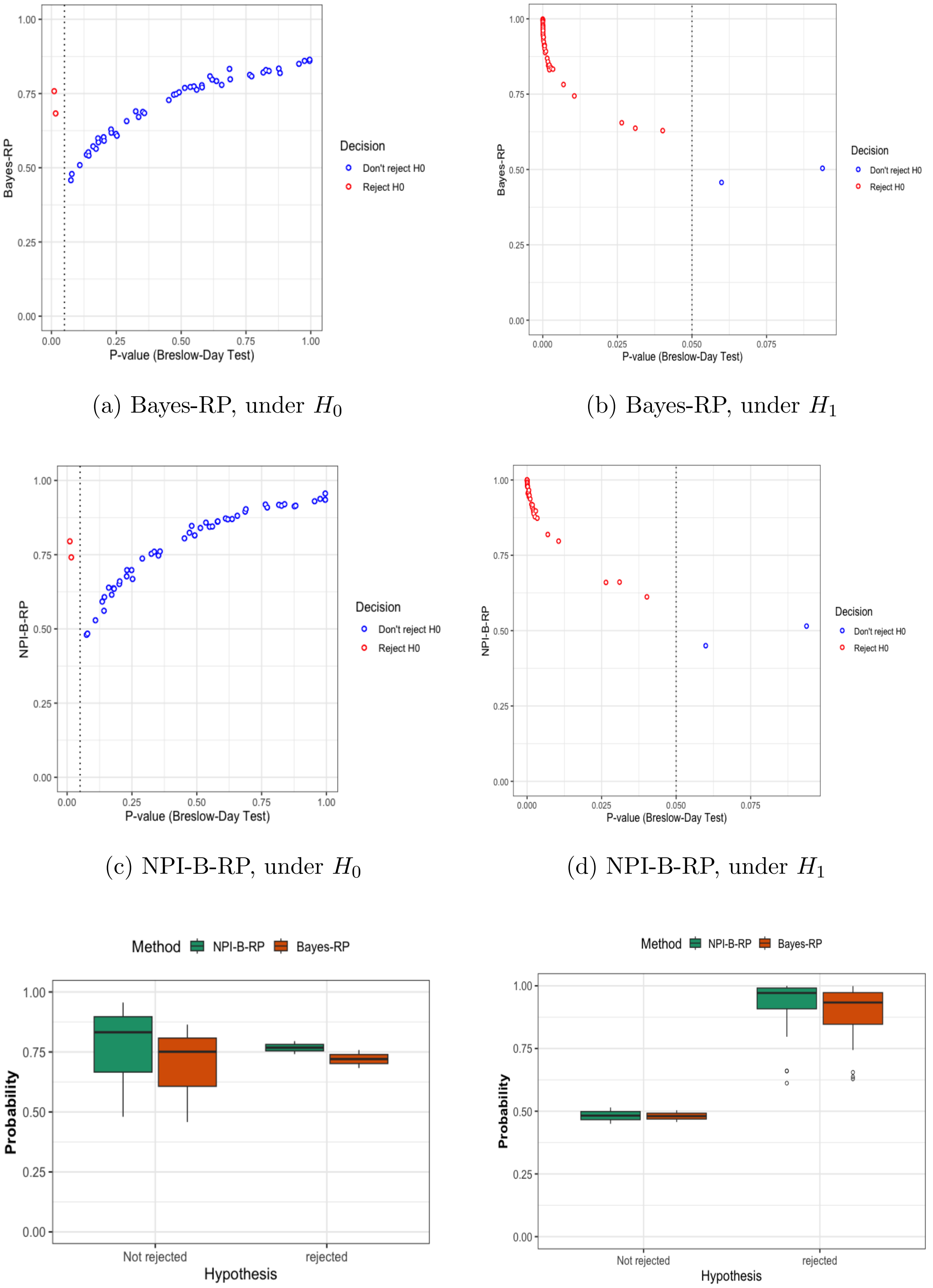
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.4: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Breslow-Day test, Scenario 1 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

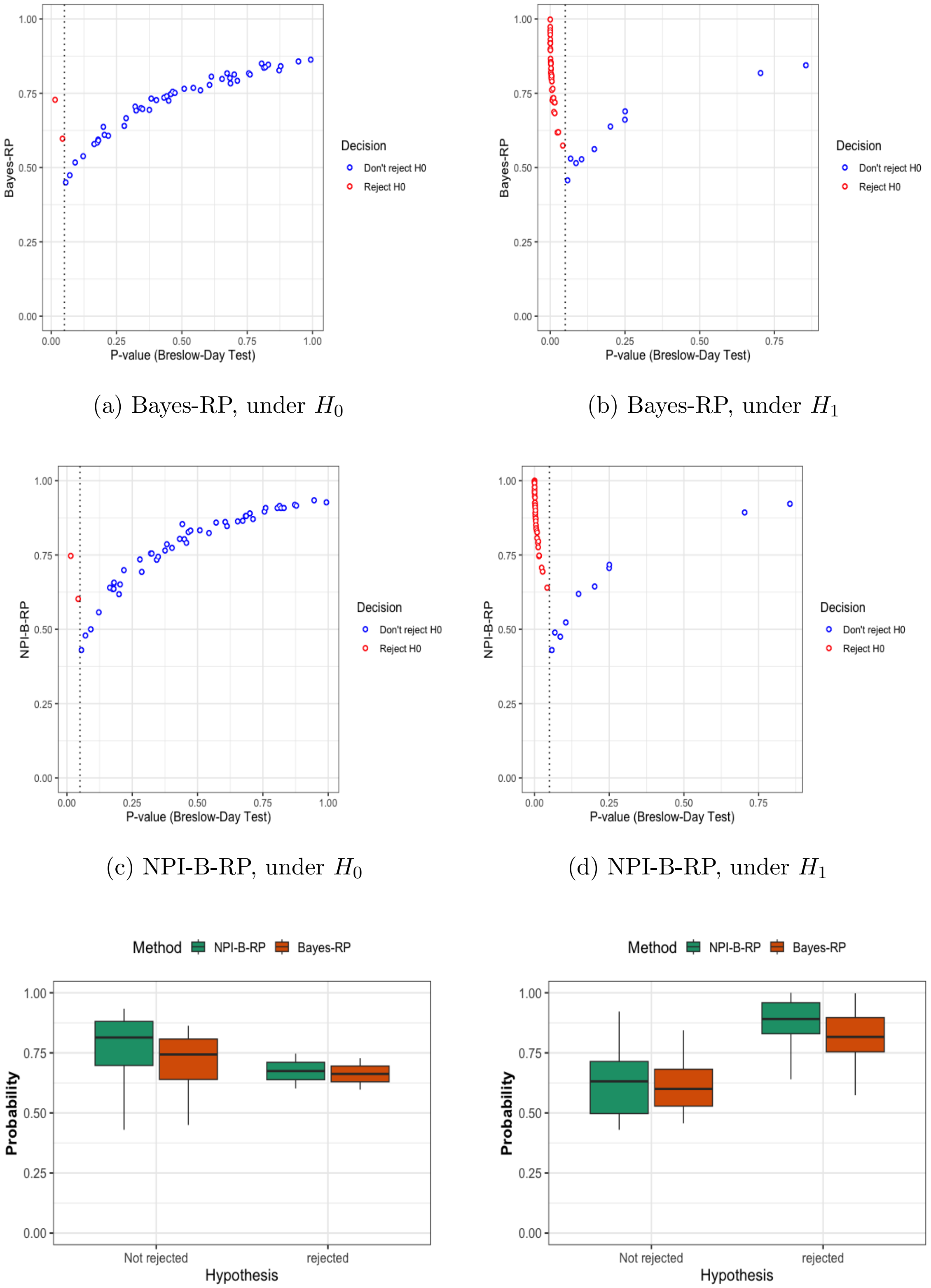
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.5: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Breslow-Day test, Scenario 2 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

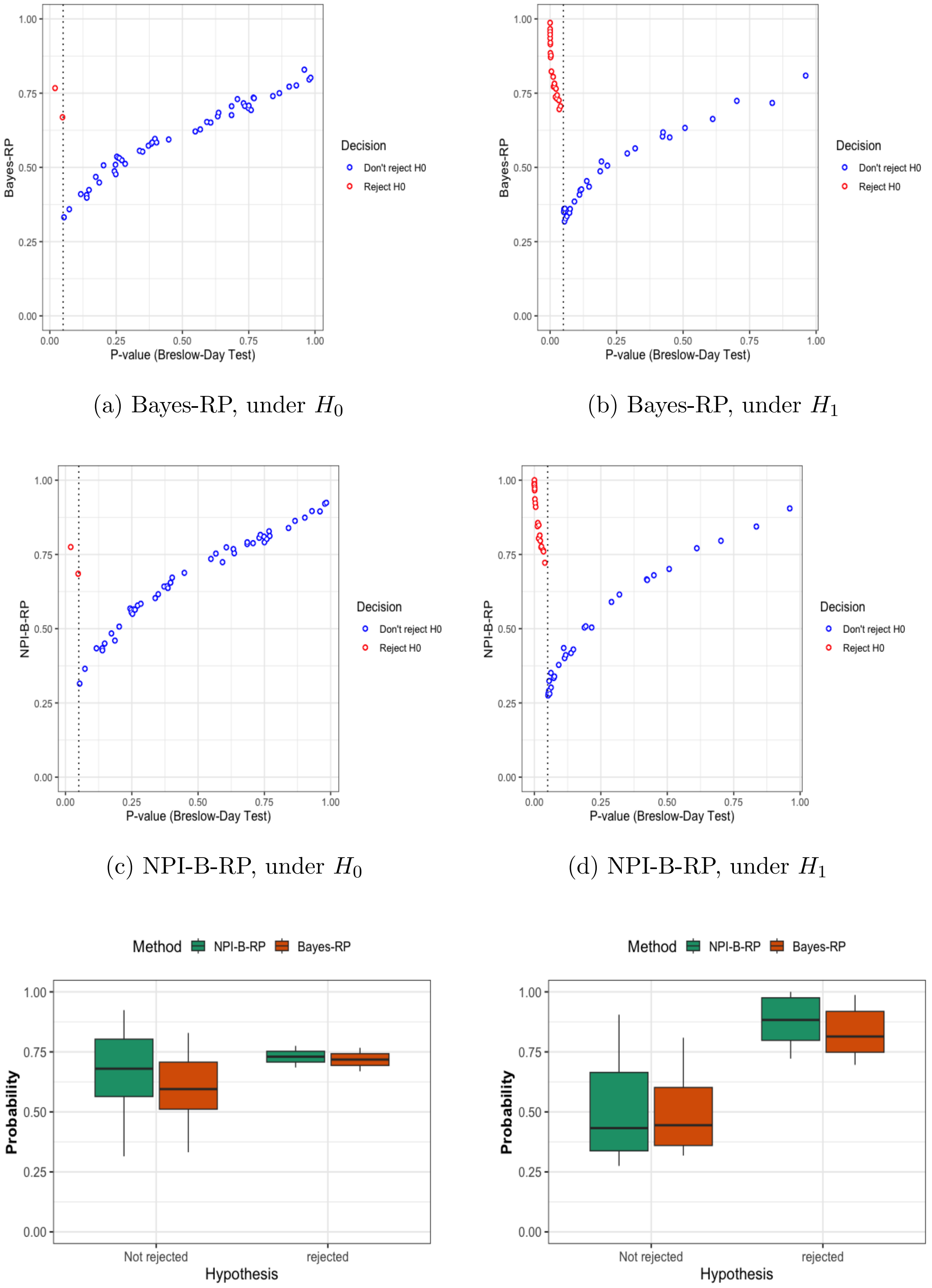
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.6: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Breslow-Day test, Scenario 3 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

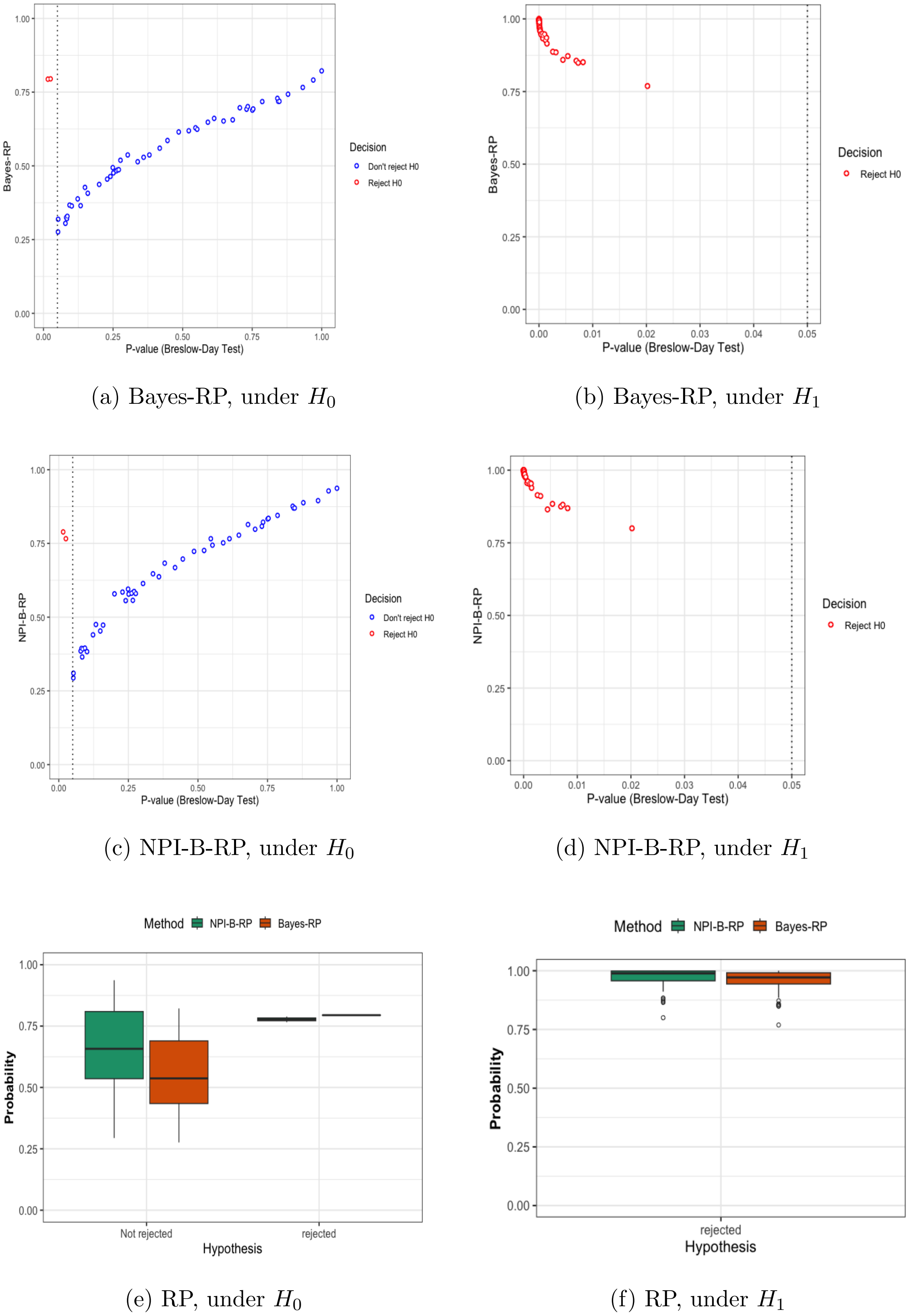
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.7: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Breslow-Day test, Scenario 1 for $K = 5$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

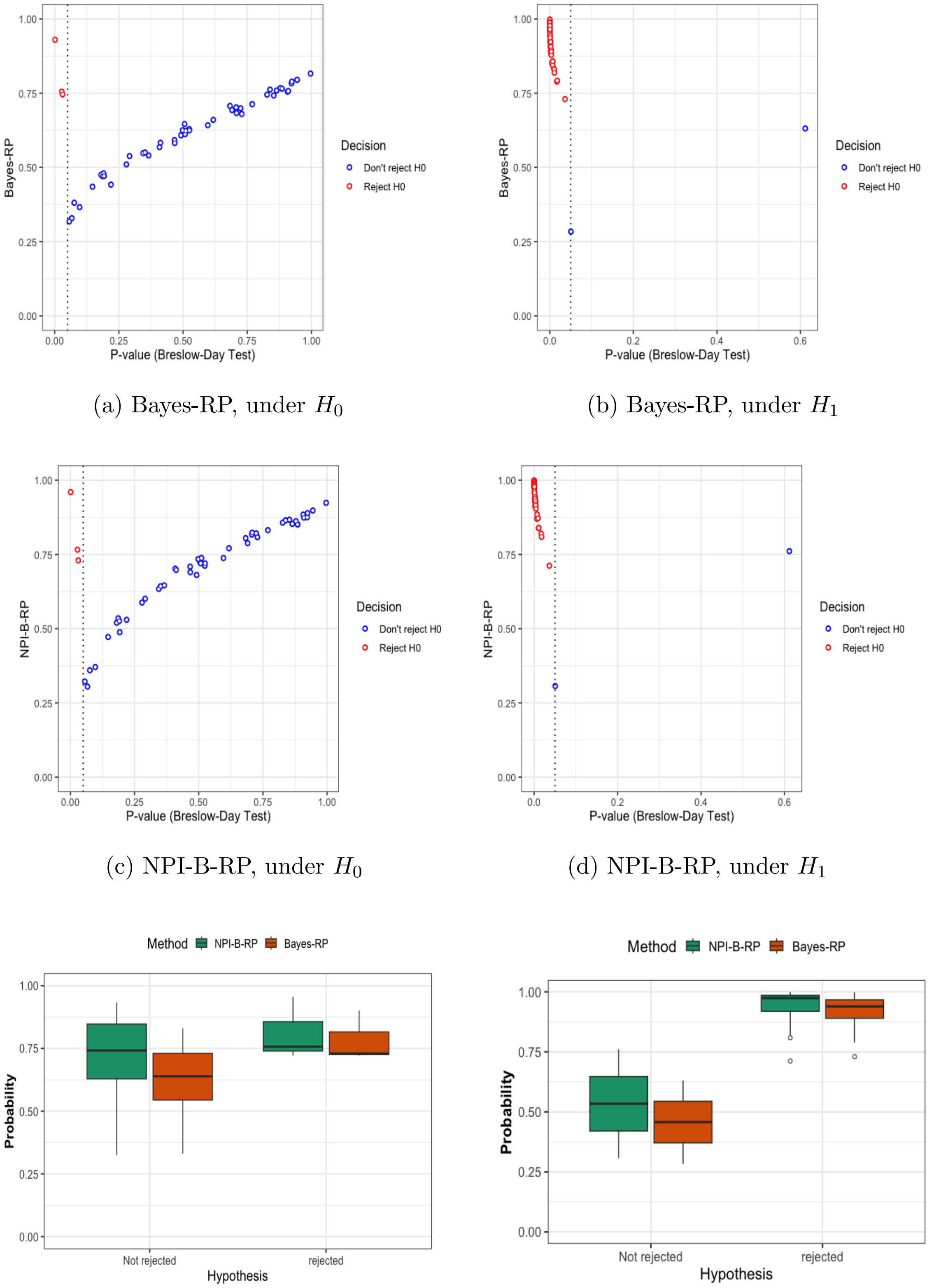
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.8: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Breslow-Day test, Scenario 2 for $K = 5$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.9: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Breslow-Day test, Scenario 3 for $K = 5$.

samples $B = 1000$ and posterior predictive samples $pp = 1000$. Our study will focus on cases where $K = 3$ and $K = 5$. As a result, the null hypothesis is generated using odds ratios $H_0 : \theta_1^{XY} = \theta_2^{XY} = \cdots = \theta_k^{XY}$, and the alternative hypothesis is generated using odds ratios H_1 : At least one of θ_k^{XY} is different from the others.

In the first case with $K = 3$, there are three suggested scenarios. Note that all simulated data sets are the same as the Breslow-Day test presented in Section 6.3. For scenario 1, Figure 6.10shows the RP values using NPI-B and Bayesian methods for RP, with boxplots representing rejections and non-rejections. Similarly, Figure 6.11 displays the RP values and boxplots for Scenario 2, and Figure 6.12 shows the same for Scenario 3. At a significance level of 0.05 and a degree of freedom of 2, the critical value is 5.99. We observe similar results as for the Woolf test with the Breslow-Day test.When the test statistic moves away from the threshold, RP tends to increase in both methods. As the sample size increases, the variability of both methods decreases, and the power of the test increases. When the p-value is 0.5 in all scenarios, NPI-B-RP is higher in non-rejection cases than Bayes-RP. Specifically, NPI-B-RP reaches values of 0.75 or higher, while Bayes-RP only approaches 0.75.

The second case with $K = 5$ suggests three scenarios. A critical value of 9.488 can be calculated when the significance level is 0.05 and the degrees of freedom are 4. For Scenario 1, Figure 6.13 shows the RP values calculated using NPI-B and Bayesian methods for RP, with boxplots representing rejections and non-rejections. Similarly, Figure 6.14 and Figure 6.15 display the RP values and boxplots for Scenarios 2 and 3, respectively. When the p-value is 0.75 with $K = 5$ in all scenarios, NPI-B-RP shows a higher in non-rejection cases compared to Bayes-RP. In particular, NPI-B-RP is higher than 0.75, while Bayes-RP approaches 0.75. In comparison, with $K = 3$ in all scenarios with the exact p-value, NPI-B-RP consistently achieves values greater than 0.85. At the same time, Bayes-RP approaches 0.85.

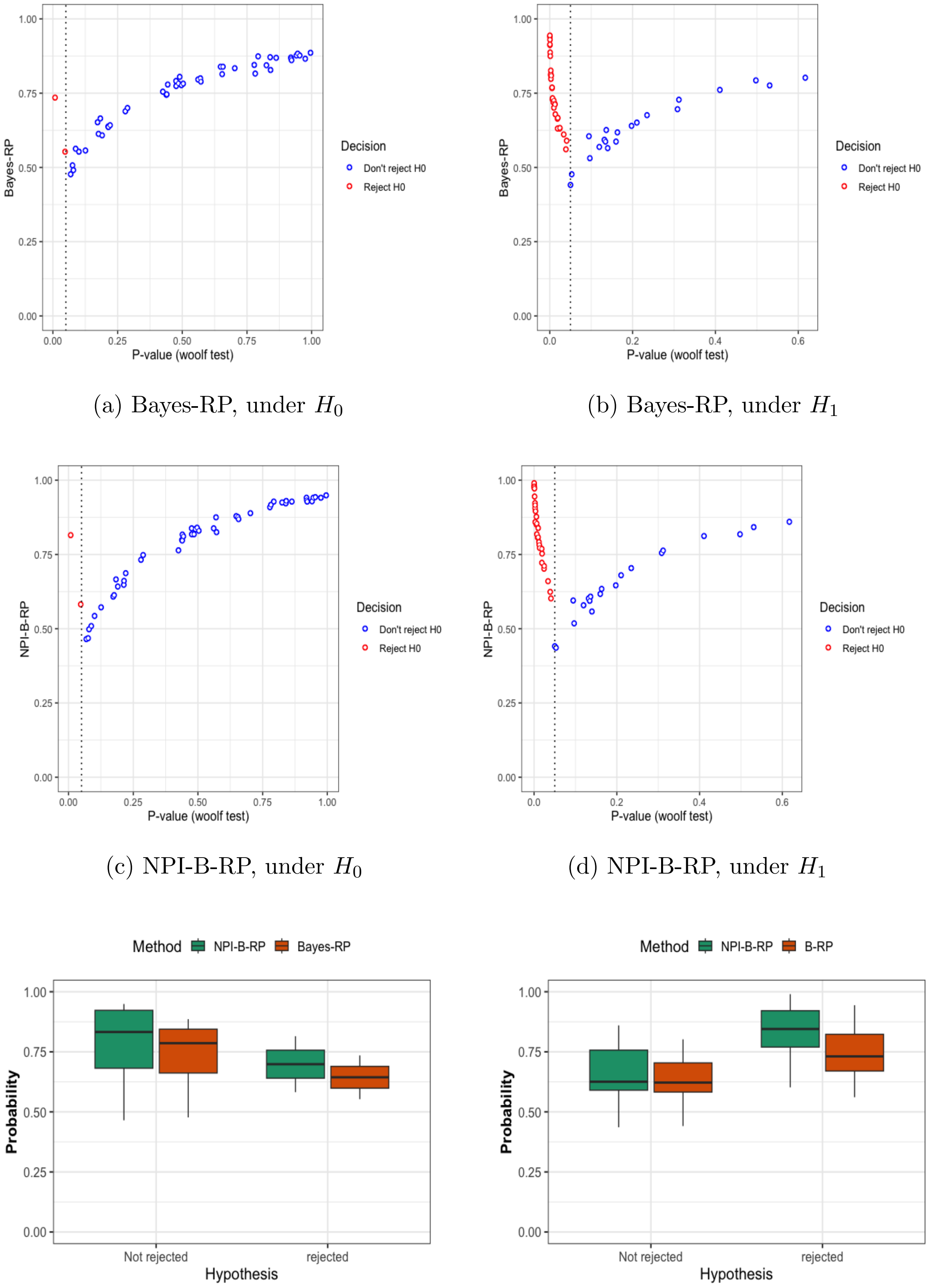
We observed that increasing the number of K tables impacts the patterns of RP values. With more K tables, both methods tend to show higher RP values in rejection cases (red cases) than in non-rejection cases (blue cases) when the test statistic is close to the threshold. The test statistic for the Woolf test is computed using Equation (6.2.5). In multiple tables, the Woolf test examines the consistency of odds ratios across all K tables. The overall test statistic is calculated by summing the individual test statistics from each of the K tables. As the number of tables increases, these combined test statistics result in a larger overall test statistic, making it more likely to exceed the critical value and reject the null hypothesis H_0.

As a result, the RP values of both methods tend to be lower in non-rejection cases and higher in rejection cases. This pattern occurs because the larger combined test statistics make it more likely for the overall test statistic to exceed the critical value, thus increasing the probability of rejecting H_0.

In this section, we compared the reproducibility of the Breslow-Day and Woolf tests using NPI-B and Bayesian methods. The hypothesis for the Breslow-Day test is different from that of the MH test. H_0 and H_1 are kept consistent in both the Woolf and Breslow-Day tests, and the results are compared between them. Note that all simulated data sets are the same for all tests. The Breslow-Day test produces similar reproducibility results to the Woolf test. This pattern is observed in both rejection and non-rejection cases, where NPI-B-RP performs better than Bayes-RP.

6.6 Concluding remarks

In this chapter, NPI-B and Bayesian methods are used to estimate RP for statistical hypothesis tests based on multiple 2×2 tables. Through simulation studies, we also compare a Bayesian method with the NPI-B for testing reproducibility. Bayesian and NPI-B are considered for reproducibility of the Mantel-Haenszel, Breslow-Day and Woolf tests. In the Mantel-Haenszel test with different K tables, as sample sizes increase, both rejection and non-rejection RP values tend to converge towards or exceed 0.5 when p-values are close to the threshold of 0.05. This occurs because the test becomes more powerful with larger samples, leading to decreased variability in RP values. The reproducibility of the Breslow-Day and Woolf tests decreases as the number of K tables increases, impacting the patterns of RP values. With more K tables, both methods show higher RP values in rejection cases (red cases) compared to non-rejection cases (blue cases) when the test statistic is close to the threshold of 0.05. In addition, RP values tend to increase as the p-value moves away from a



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

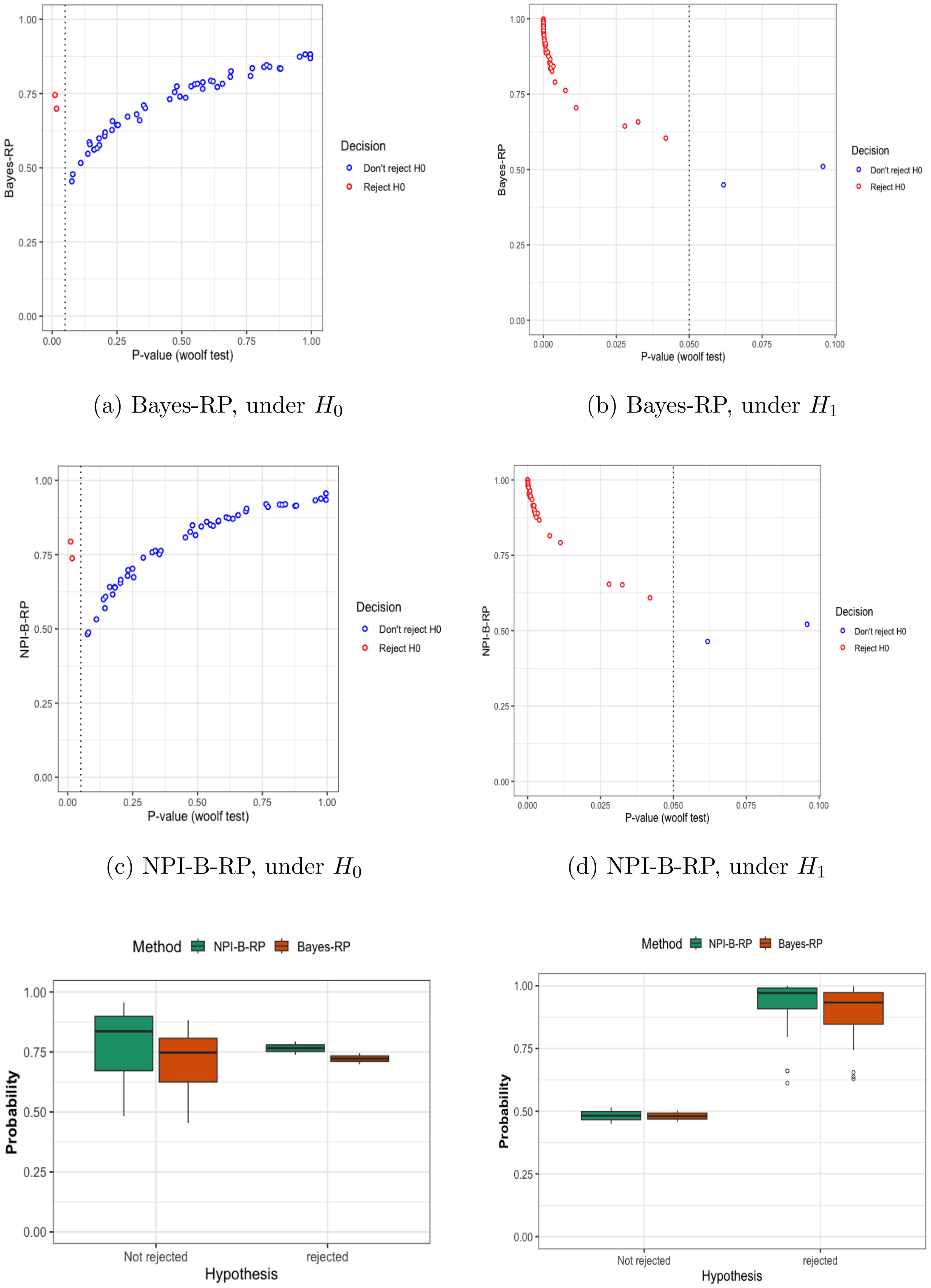
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.10: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Woolf Test, Scenario 1 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

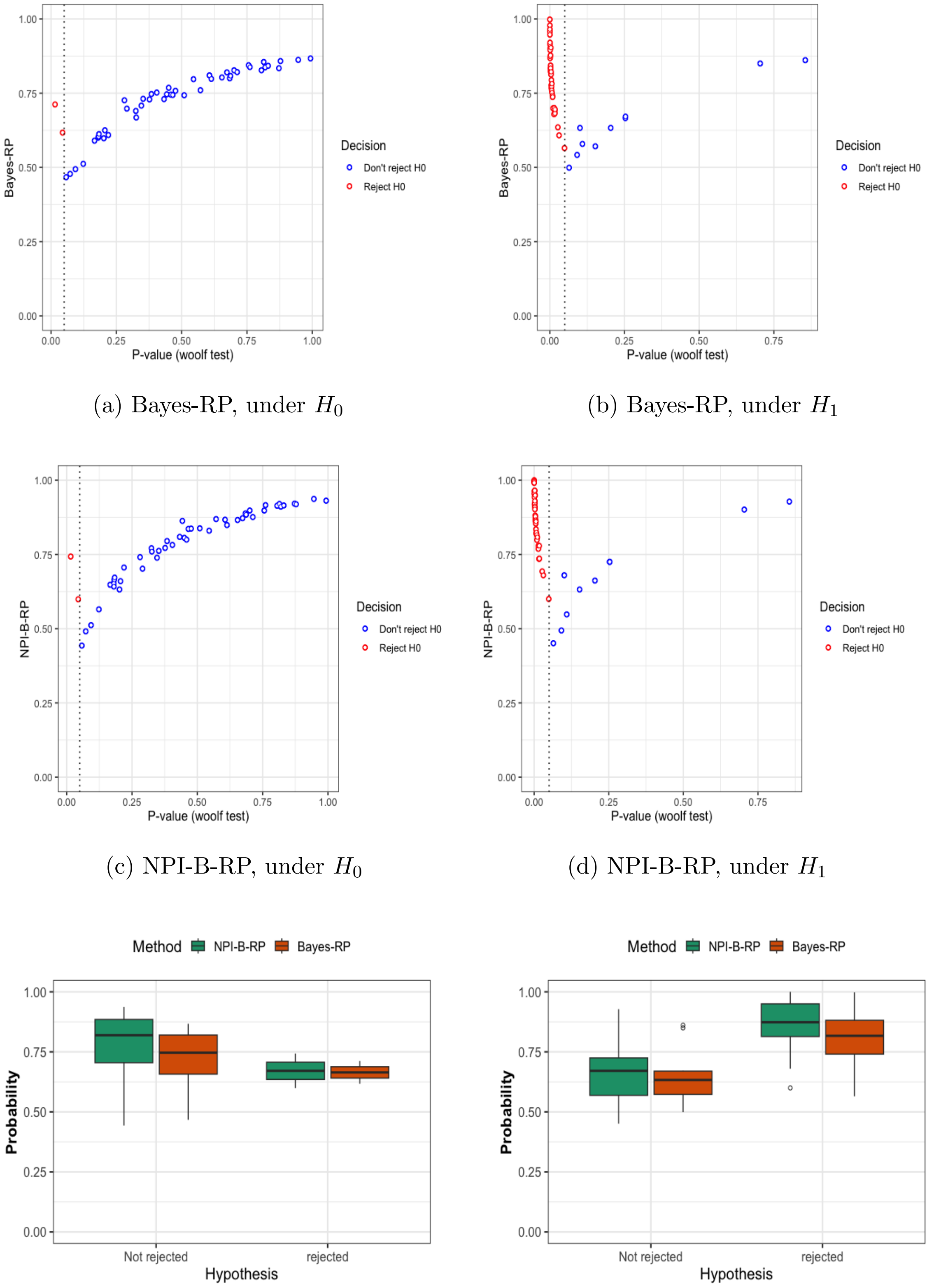
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.11: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Woolf Test, Scenario 2 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

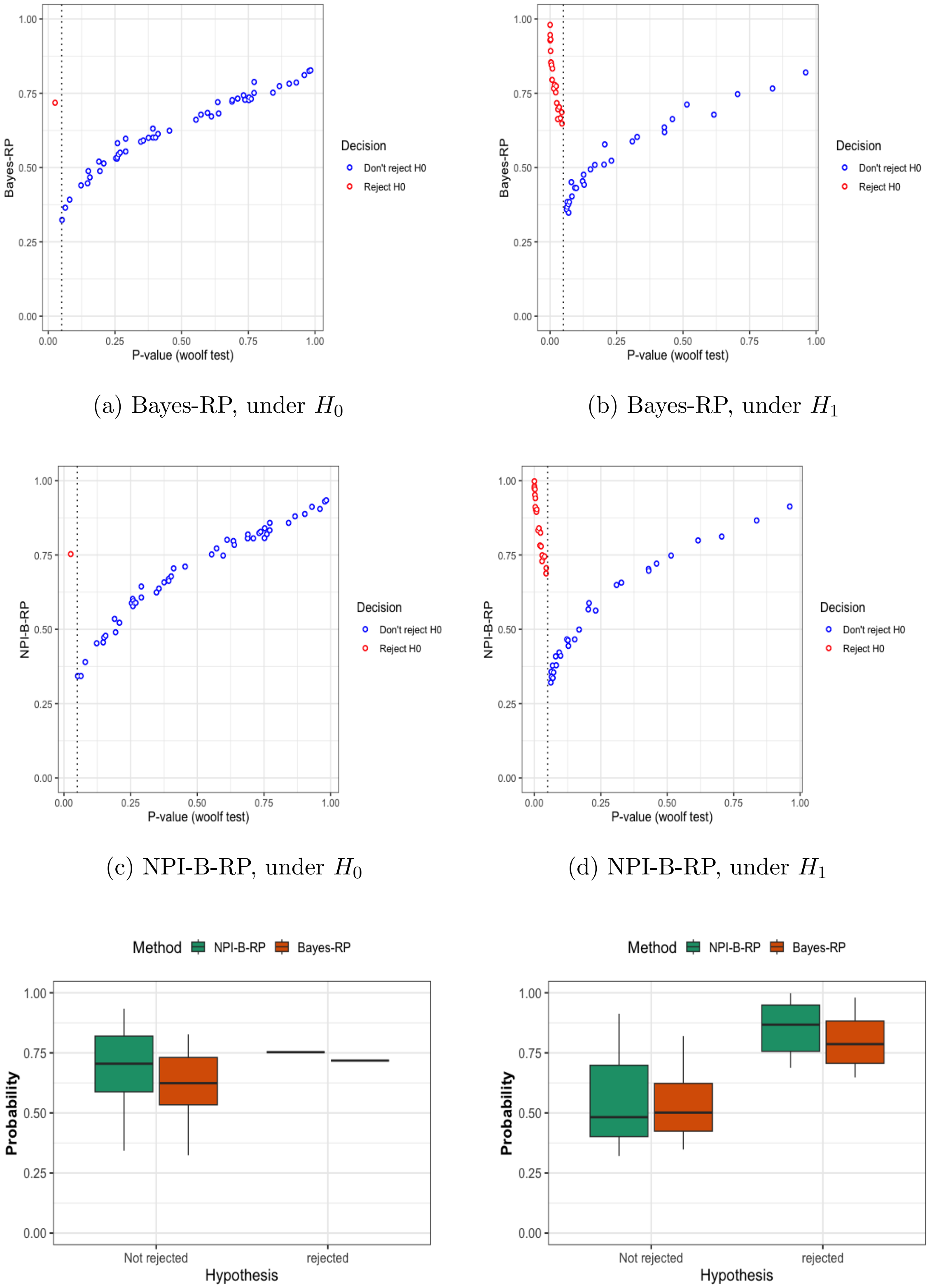
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.12: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Woolf Test, Scenario 3 for $K = 3$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

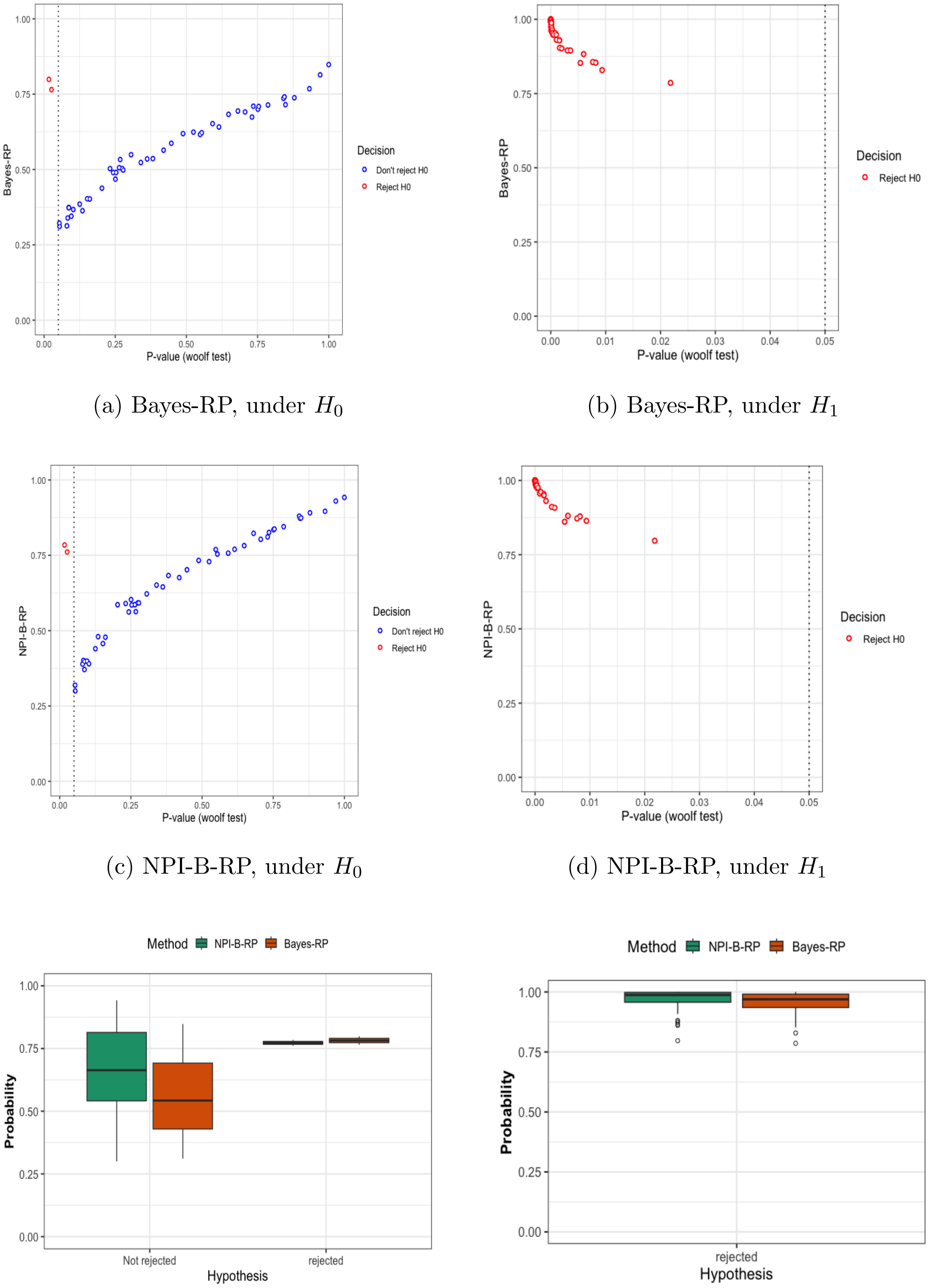
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.13: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Woolf Test, Scenario 1 for $K = 5$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

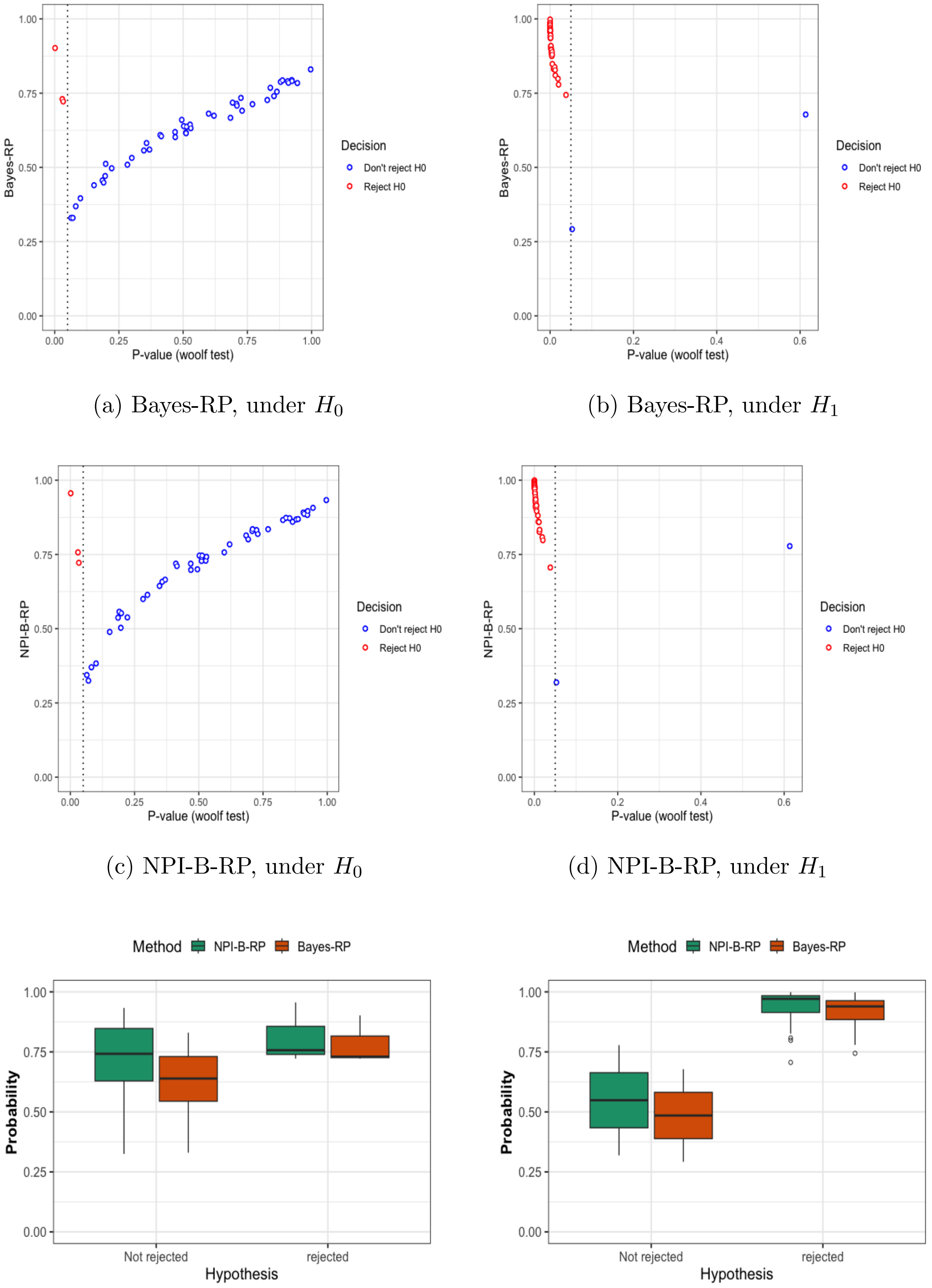
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.14: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Woolf Test, Scenario 2 for $K = 5$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure 6.15: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Woolf Test, Scenario 3 for $K = 5$.

significance level of 0.05.

Chapter 7

Conclusions

In this chapter, we present a summary of the main results of this thesis and conclude with some recommendations for future research. In this thesis, we introduce the generalisation of NPI for the 2 × 2 table and the reproducibility of hypotheses tests based on the 2 × 2 table and multiple 2 × 2 tables.

In Chapter 3, NPI was introduced for circular data as a method for inference based on 2 × 2 table data. The NPI lower and upper probabilities are computed for one future observation based on the assumption $\textcircled{A}_{(n)}$ and for multiple future observations based on the assumption $\textcircled{A}_{(\bullet)}$. For single and multiple future observations, the NPI lower probabilities can be derived results in exact formulas. On the other hand while it is trivial to derive the upper probability for one future observation and it is not easy to derive the upper probability considering multiple future observations, so we are proposing some approximations. Therefore, an algorithm has been proposed to find approximations using sampling of ordering method for estimating upper. For multiple future observations, the NPI lower probabilities are approximated and compared to the results computed by the exact formula. Additionally, the NPI-B was introduced as a computational version for the circular data. It is used for inferences in our NPI method with 2 × 2 table data.

In Chapter 4, we approximated the reproducibility for tests of independence, Fisher's exact test, and McNemar's test using two methods NPI bootstrap and NPI sampling of orderings. Test reproducibility aligns well with the explicit predictive nature of NPI and NPI-B. It is the probability that the same test outcome would be obtained if a test were repeated under identical circumstances with the same sample size. In recent years, interest in reproducibility (RP) has grown significantly due to its critical role in assessing the practical relevance of test results. The RP serves as a measure of the reliability of statistical hypothesis test outcomes, making it a fundamental concept in scientific methodology. It provides researchers with confidence and clarity about the robustness and validity of their findings. Through simulation studies, we have studied the reproducibility of tests using the NPI sampling of orderings and NPI-B methods. When the sample size is increased, the variability in the RP values decreases, due to the power of the test. The bootstrap approach to predicting RP avoids the difficulty of determining the lower and upper boundaries in NPI-RP. In NPI-B-RP, we present the RP as a point estimate rather than the lower and upper values. We also compared the reproducibility of the chi-square test of independence, the likelihood ratio test of independence, and McNemar's test in the simulations, and found that the results for these tests were similar.

Chapter 5 covered the Bayesian method for evaluating the reproducibility of statistical hypotheses based on 2×2 table. We used both the Bayesian method and NPI bootstrap to estimate reproducibility for tests of independence, Fisher's exact test, and McNemar's test. We conducted simulations for Bayesian Inference and NPI-B, comparing both methods for the reproducibility of a variety of tests. Test reproducibility is more of a prediction issue than an estimation issue, aligning well with the explicit predictive nature of both Bayesian and NPI-B methods, which consider future observations. As sample sizes increase, the RP values based on Bayesian and NPI-bootstrap methods tend to converge towards 0.5 for both rejection and non-rejection when p-values are close to the significance level of 0.05. This convergence is primarily due to the more powerful test that results from a larger sample size, leading to decreased variability in RP values.

Finally, Chapter 6 presented RP estimation for statistical hypothesis tests based on multiple 2×2 tables, utilising both NPI-B and Bayesian methods. We conducted simulation studies and compared the reproducibility of both method for Mantel-Haenszel, Breslow-Day, and Woolf tests. The reproducibility of The Mantel-Haenszel test values for both methods show less variability with larger sample sizes due to the increased power of the test. With an increase in K tables for the Breslow-Day and Woolf tests, both methods tend to yield higher values in rejection cases than in non-rejection cases when the test statistic is close to the threshold 0.05. Additionally, RP values tend to increase as the p-value moves away from the threshold.

In this thesis, we have studied the NPI method for 2×2 tables. This work opens up new opportunities for future research, including the potential application of NPI to $r \times c$ tables. We have also presented both NPI and Bayesian methods to determine reproducibility of hypothesis tests based on 2×2 tables. This work could be extended for reproducibility of hypothesis tests based on $r \times c$ tables. To extend this work to $r \times c$ tables, we could adapt the NPI method to handle more rows and columns, considering the relationships between them. The main challenges would be the increased complexity of calculations and understanding how the larger table structure affects reproducibility. Future researchers could focus on finding simpler methods and studying these effects.

In addition, we quantified reproducibility within the Bayesian framework and inferred future observations through the posterior predictive distribution. Previous studies have explored RP based on NPI for various tests such one sample sign test, one sample signed rank test, two sample rank sum test, and the two-sample Kolmogorov-Smirnov test [11]. It would be interesting to consider RP for these tests under the Bayesian framework. We also recommend further research to compare the NPI-RP and Bayesian reproducibility for these tests.

Appendix A

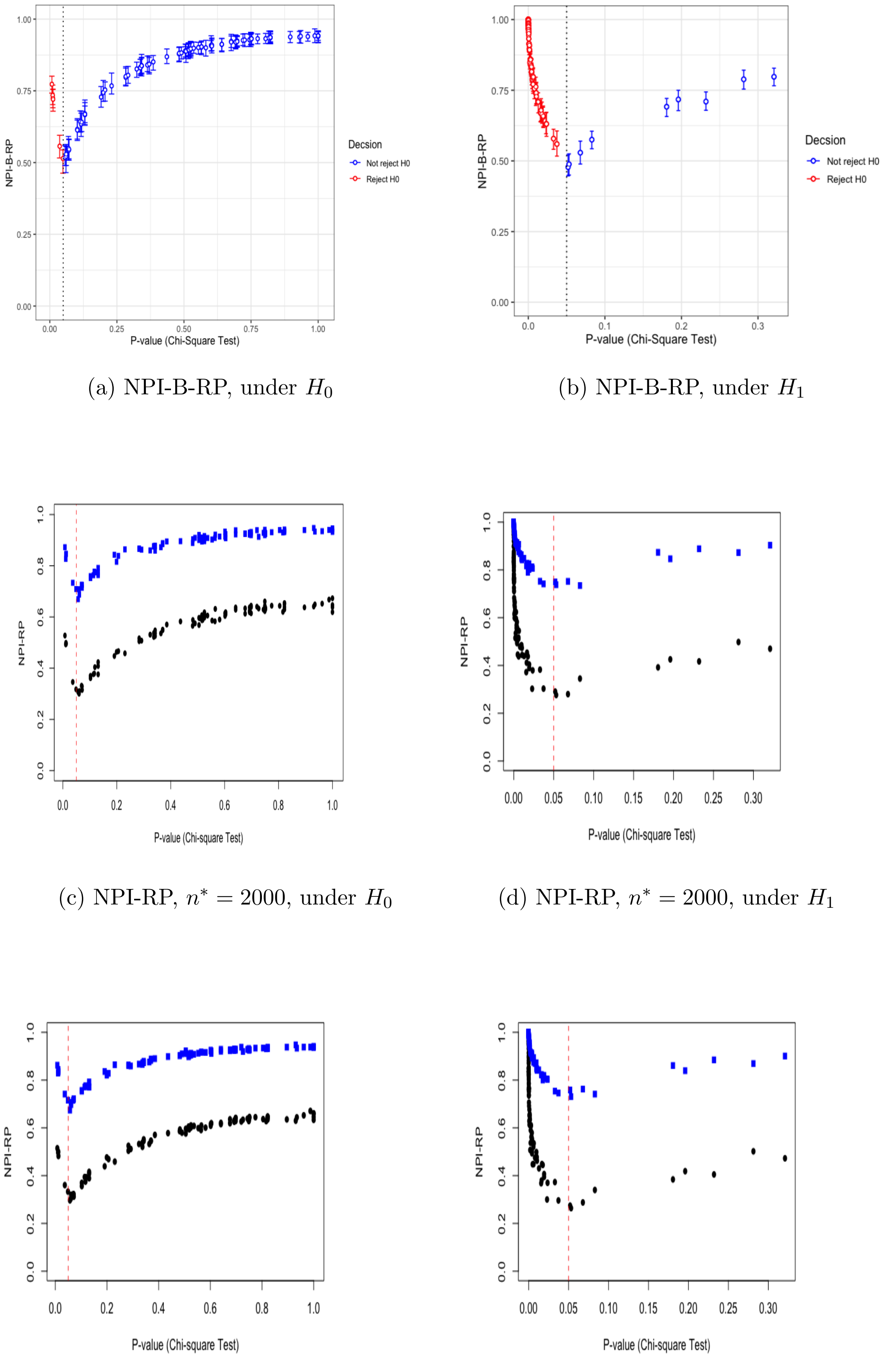
Basic Results

A.1 NPI and NPI-B-RP for test of independence

In this study, we investigate NPI-RP and NPI-B-RP for the chi-square test of independence and the likelihood ratio test of independence. We generate data from the a multinomial distribution with probabilities (0.25,0.25,0.25,0.25) under the null hypothesis and with probabilities (0.6,0.1,0.1,0.2) under the alternative hypothesis. This simulation study uses the following inputs: $n = 30, 60$ and $N = 100$ simulations.

A.2 NPI and NPI-B-RP for McNemar test

In this simulation study, the following inputs are used: $n = 30, 60$, and $N = 100$ simulations per run. Simulations are conducted to evaluate the NPI and NPI-B methods for the RP of McNemar test of RP in accordance. The simulations are performed both under H0 and H1. Under H0, data are generated from the a multinomial distribution with probabilities (0.2, 0.3, 0.3, 0.2). Under H1 data are generated from a multinomial distribution with probabilities (0.3, 0.2, 0.4, 0.1).



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

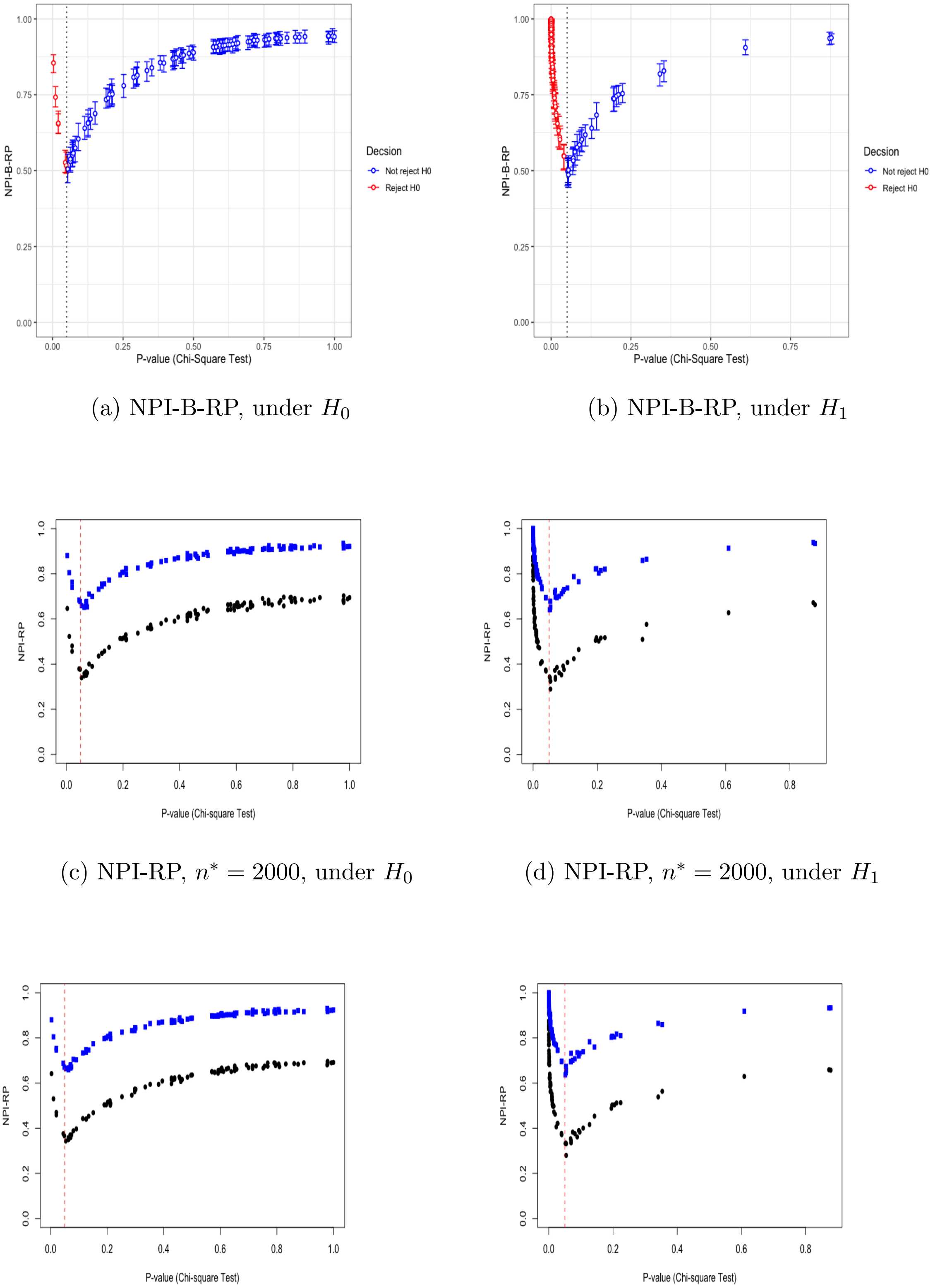
(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure A.1: Simulations under H_0 and H_1: NPI-RP and NPI-B-RP values for chi-square testing of independence, where $n = 30$.



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

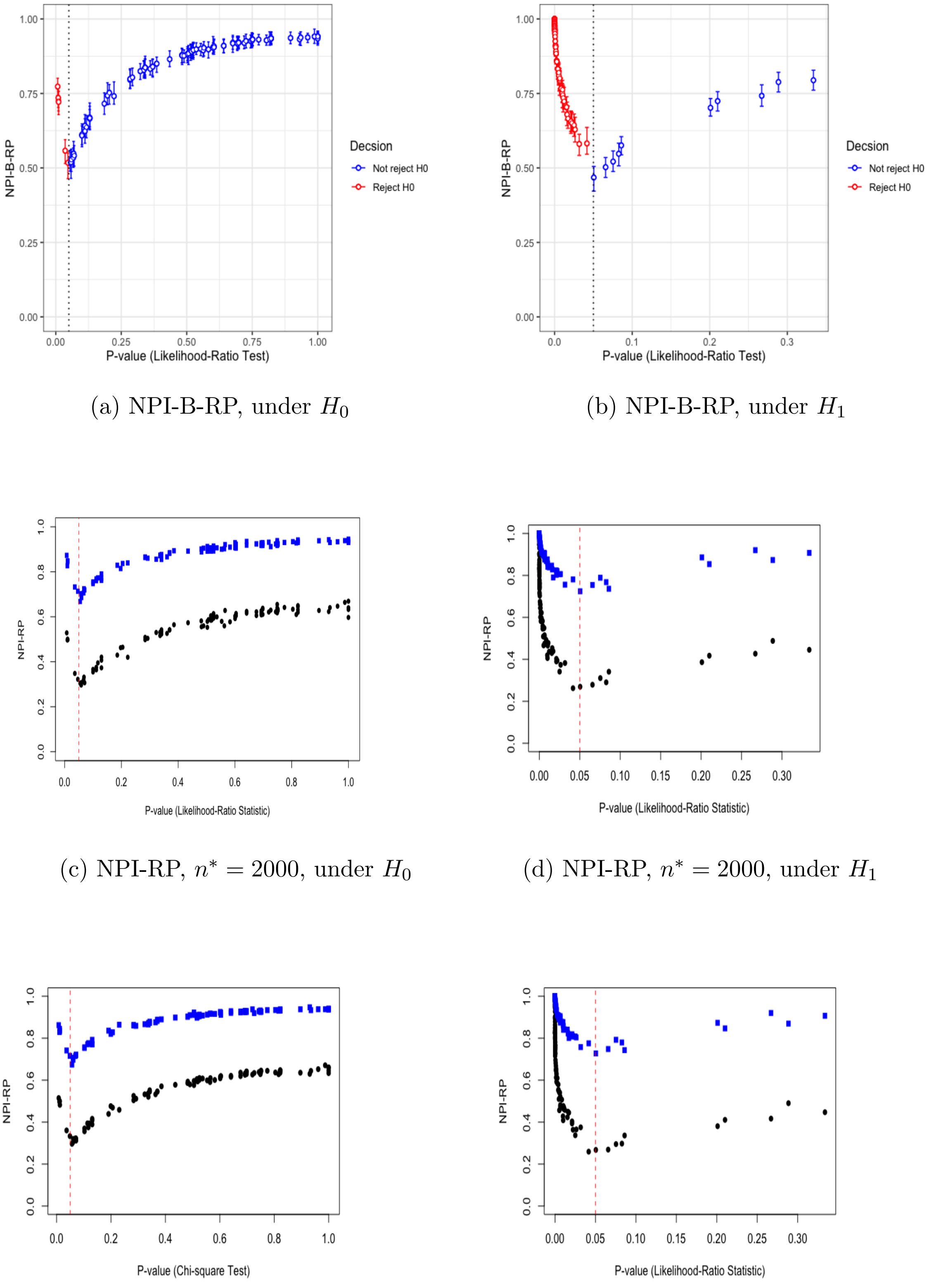
(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure A.2: Simulations under H_0 and H_1:NPI-RP and NPI-B-RP values for chi-square testing of independence, where $n = 60$.



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

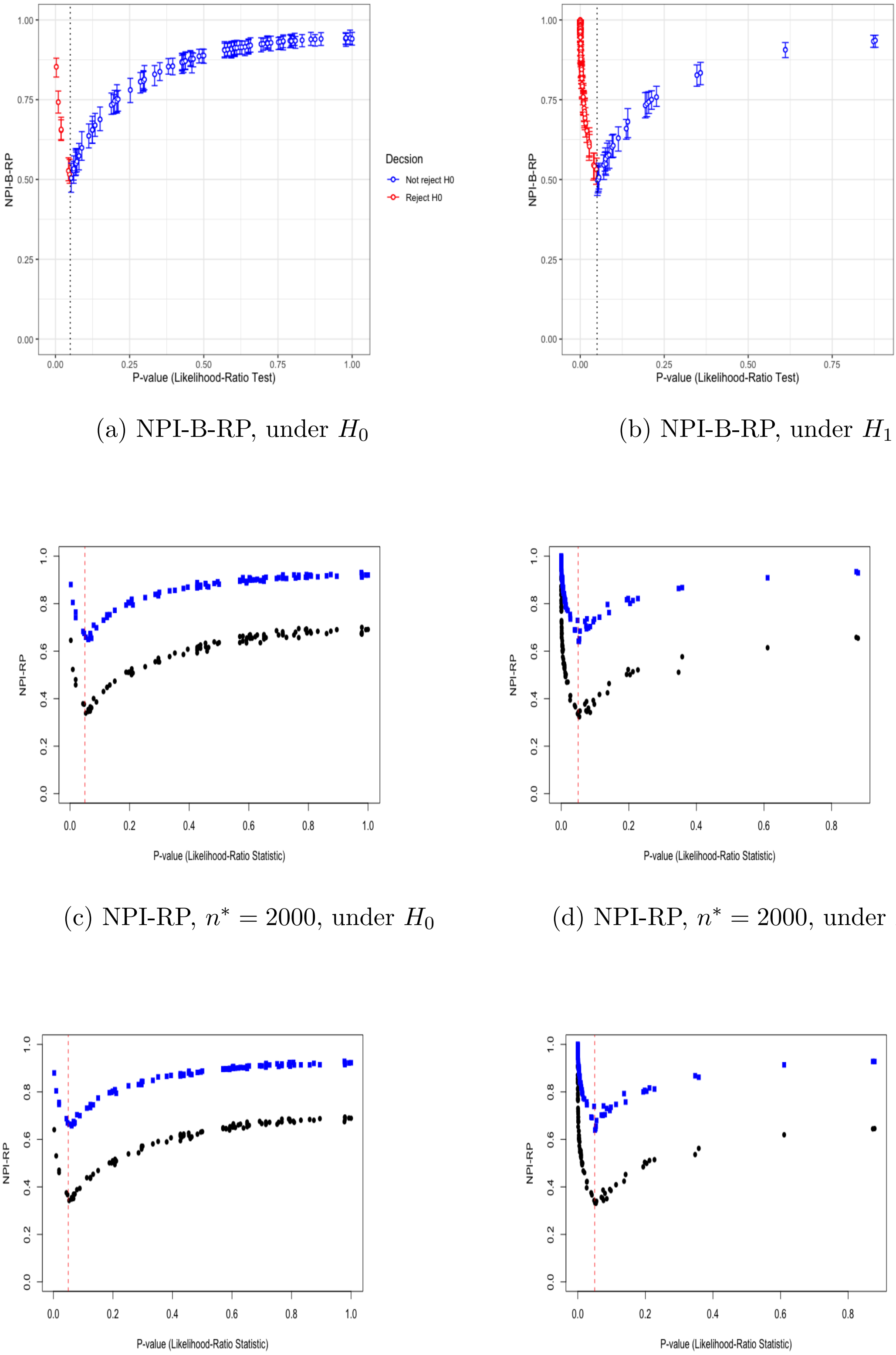
(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure A.3: Simulations under H_0 and H_1: NPI-RP and NPI-B-RP values for likelihood ratio testing of independence, where $n = 30$.



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

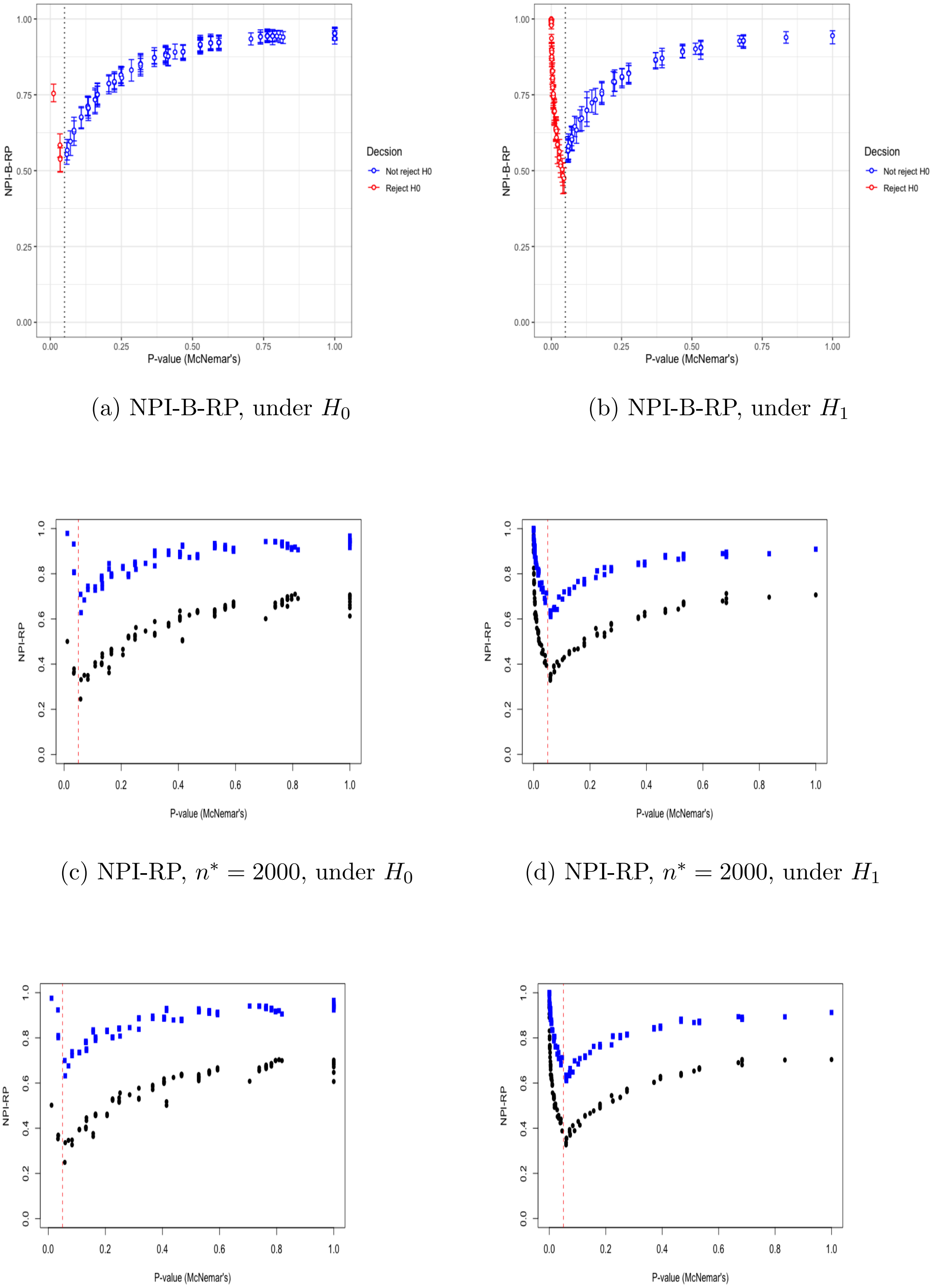
(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure A.4: Simulations under H_0 and H_1:NPI-RP and NPI-B-RP values for likelihood ratio testing of independence, where $n = 60$.



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

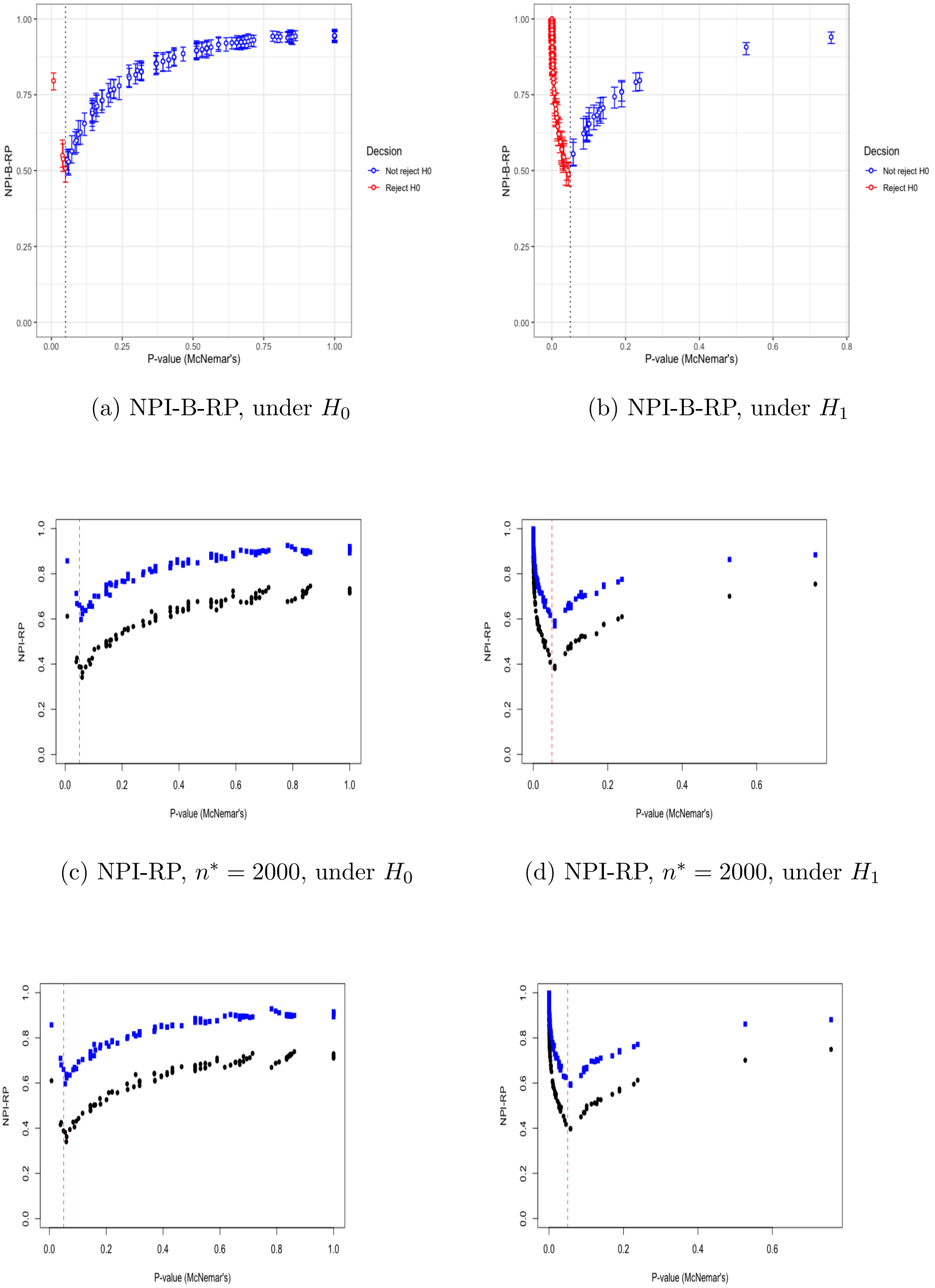
(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure A.5: Simulations under H_0 and H_1: NPI-RP and NPI-B-RP values for McNemar's test, where $n = 30$.



(a) NPI-B-RP, under H_0

(b) NPI-B-RP, under H_1

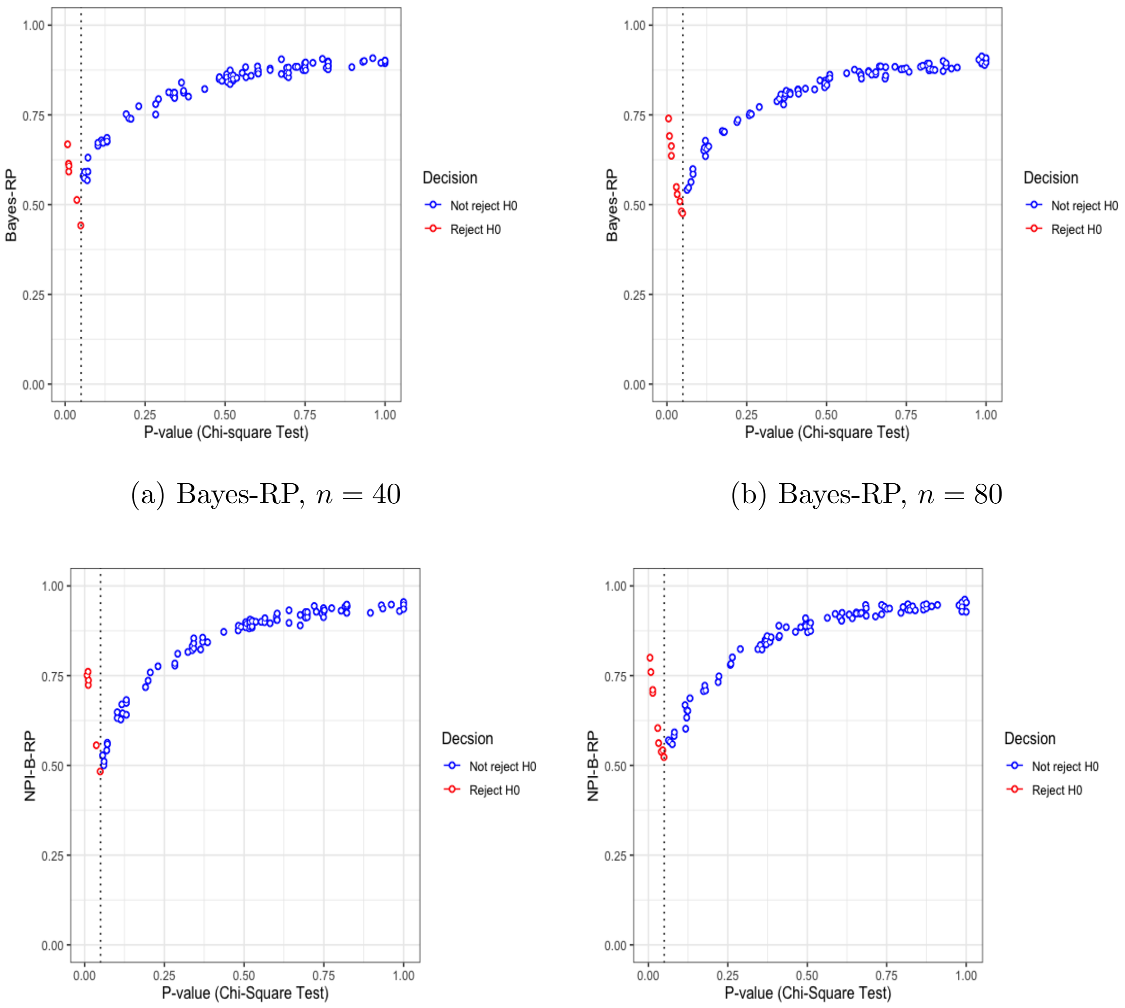
(c) NPI-RP, $n^* = 2000$, under H_0

(d) NPI-RP, $n^* = 2000$, under H_1

(e) NPI-RP, $n^* = 5000$, under H_0

(f) NPI-RP, $n^* = 5000$, under H_1

Figure A.6: Simulations under H_0 and H_1: NPI-RP and NPI-B-RP values for McNemar's test, where $n = 60$.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

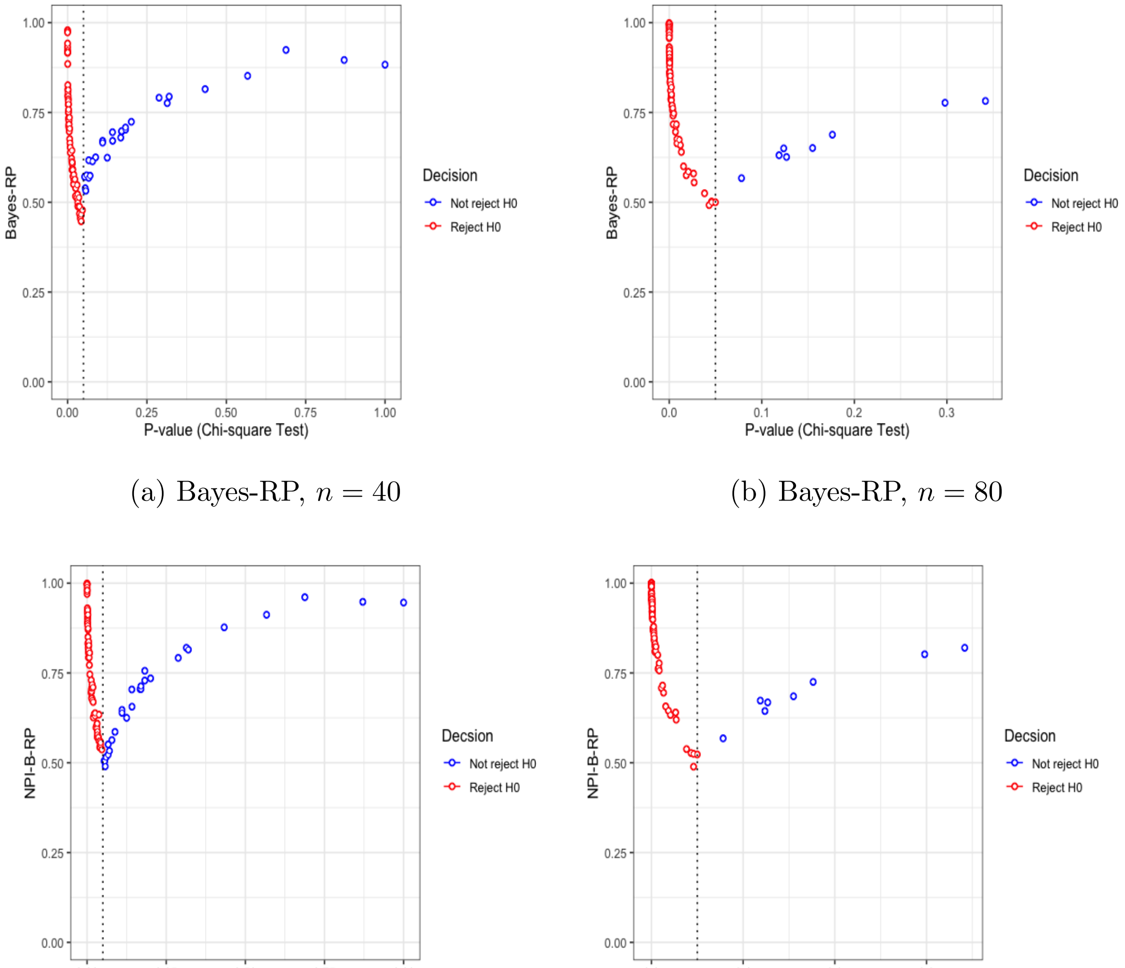
(c) NPI-B-RP, $n = 40$

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

Figure A.7: Simulations under H_0 : Bayes-RP and NPI-B-RP values for chi-square test of independence.

A.3 Bayes-RP and NPI-B-RP for chi-square test of independence and McNemar's Test

The data are generated from the multinomial distributions with probabilities (0.25, 0.25, 0.25, 0.25). The data are generated H_1 from a multinomial distribution with probabilities (0.1, 0.5, 0.2, 0.2). We simulate $N = 100$ samples of sizes $n = 40, 80$.



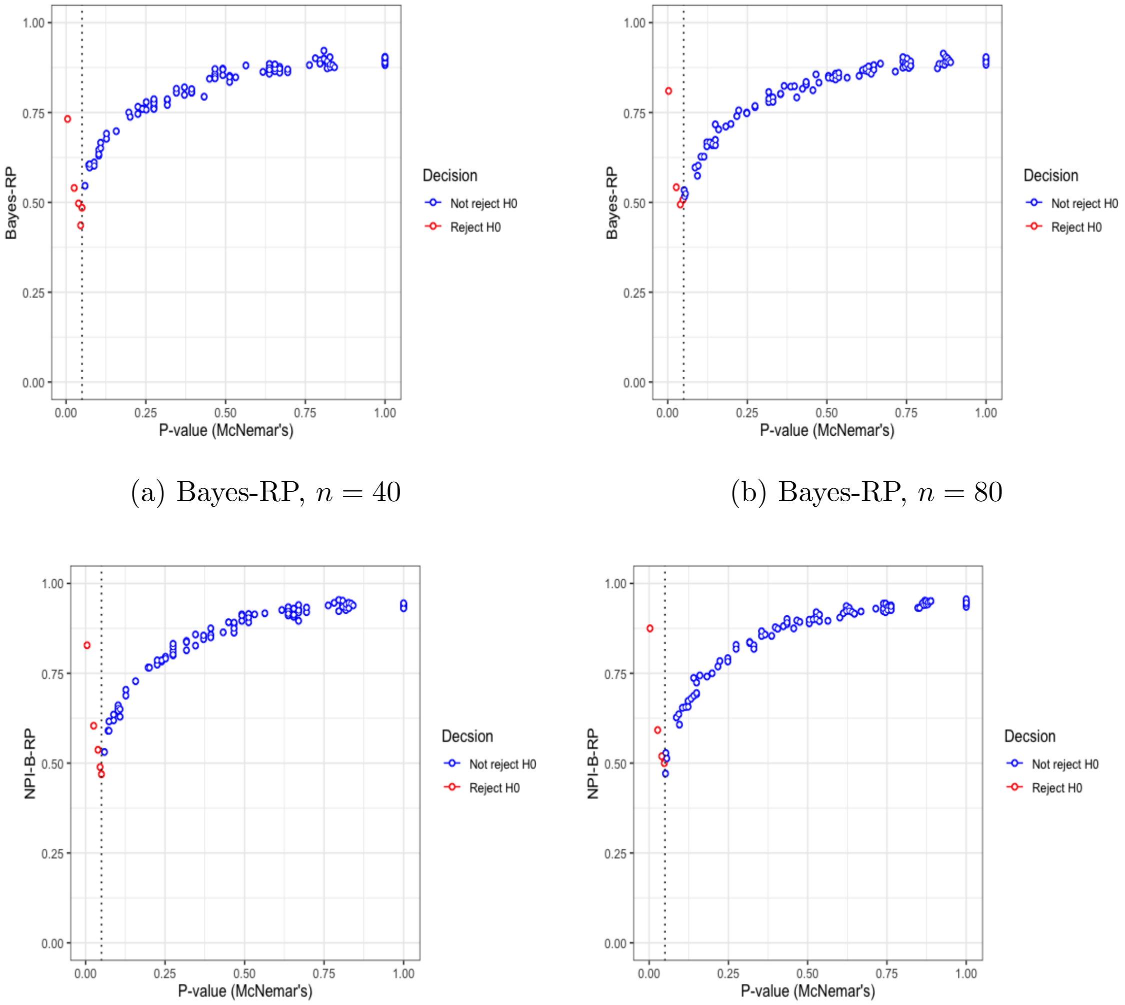
(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

(c) NPI-B-RP, $n = 40$

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

Figure A.8: Simulations under H_1 with probabilities $(0.1, 0.5, 0.2, 0.2)$: Bayes-RP and NPI-B-RP values for chi-square test of independence.



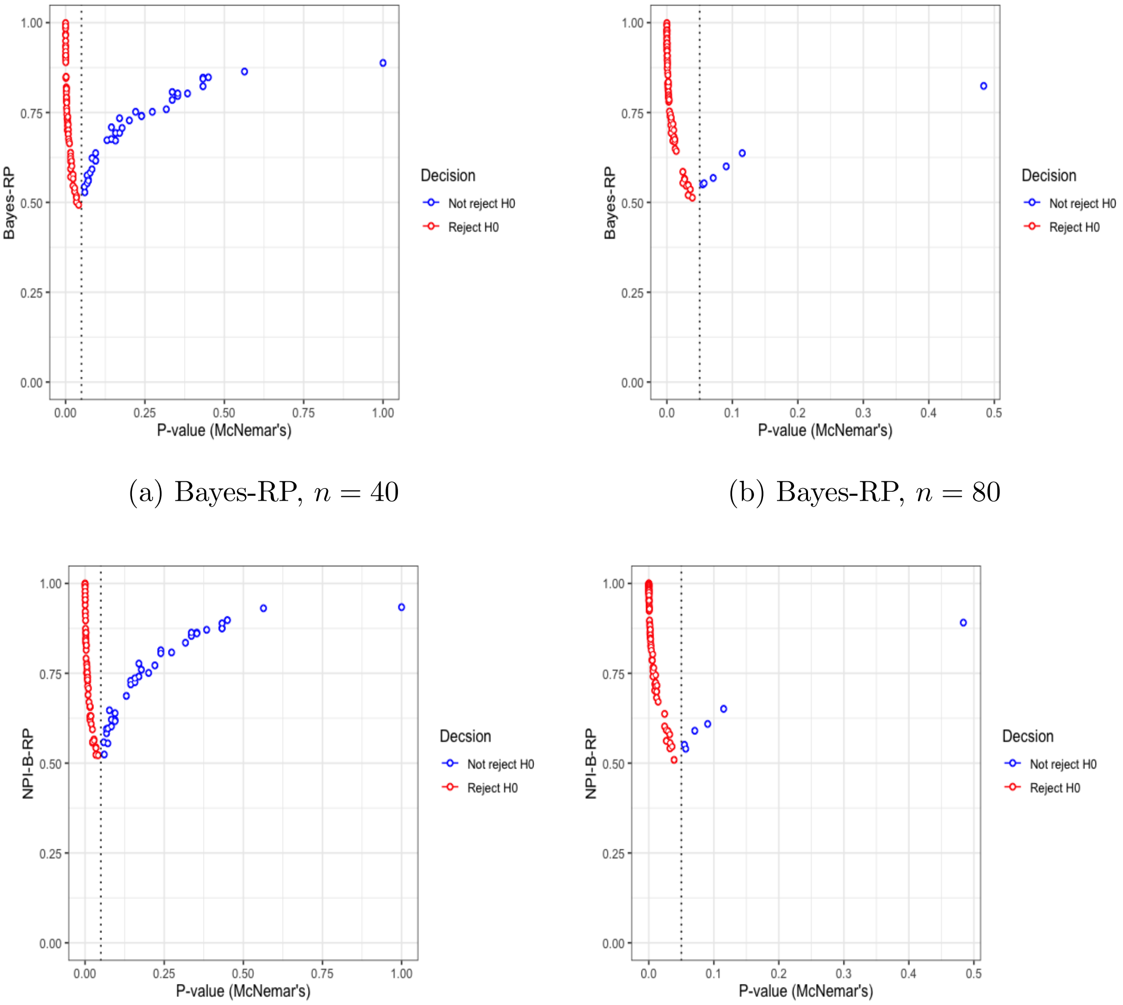
(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

(c) NPI-B-RP, $n = 40$

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

Figure A.9: Simulations under H_0 : Bayes-RP and NPI-B-RP values for McNemar's Test.



(a) Bayes-RP, $n = 40$

(b) Bayes-RP, $n = 80$

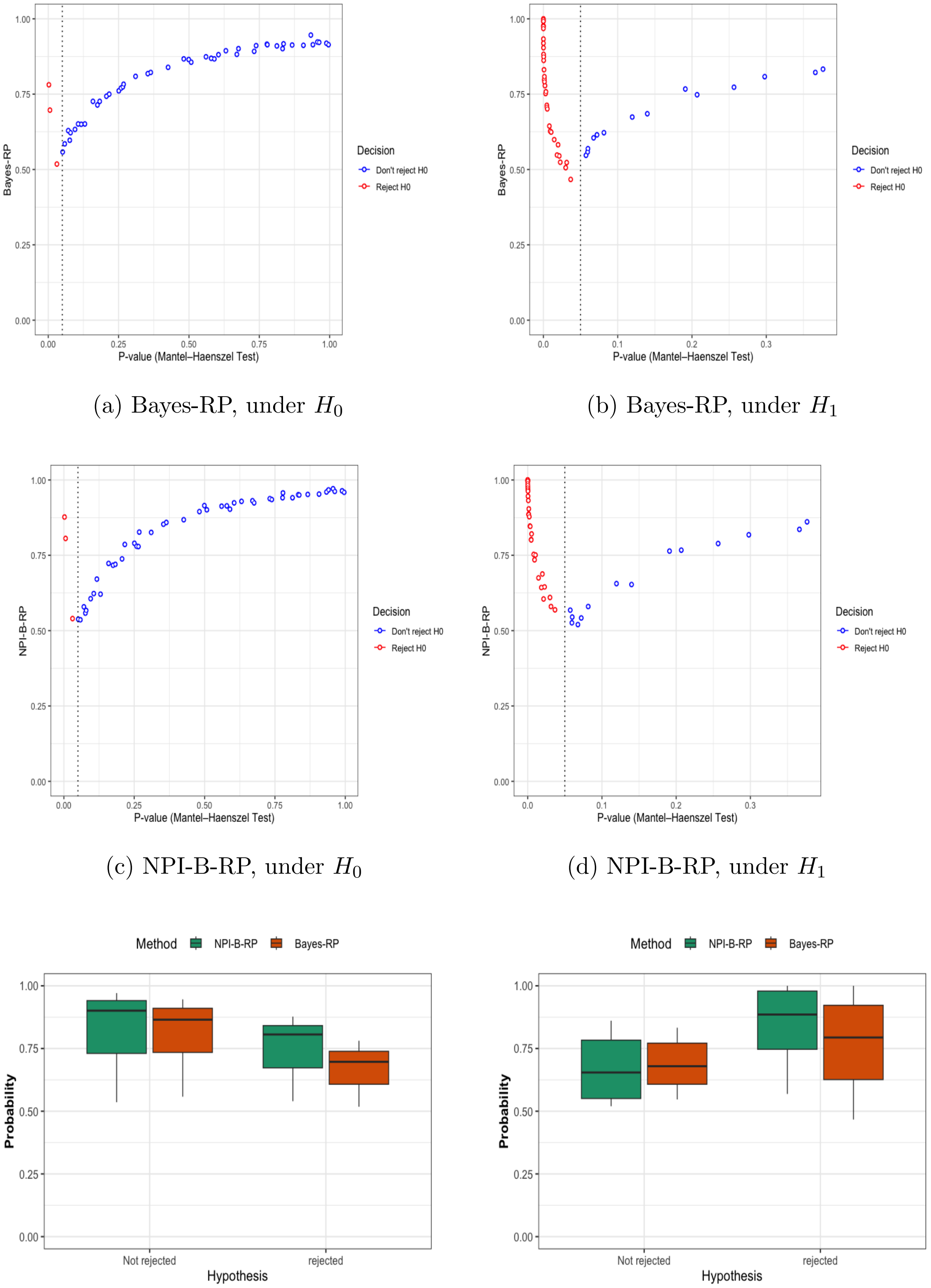
(c) NPI-B-RP, $n = 40$

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

Figure A.10: Simulations under H_1 with probabilities $(0.1, 0.5, 0.2, 0.2)$: Bayes-RP and NPI-B-RP values for McNemar's test.

A.4 Reproducibility for Mantel-Haenszel test

Based on the data generated for the second case with $K = 5$, three scenarios can be identified. According to scenario 1, the sample sizes are $n_1 = 60, n_2 = 55, n_3 = 50, n_4 = 40$, and $n_5 = 45$. As shown in Figure A.11, RP values are obtained using bootstrap and posterior predictive methods under H_0 and H_1 for the NPI-B-RP and Bayes-RP methods. A boxplot is presented for both cases of rejection and non-rejection. Sample sizes for scenario 2 are $n_1 = 200$, $n_2 = 180$, $n_3 = 160$, $n_4 = 140$, and $n_5 = 150$. Figure A.12 similarly illustrates the RP value and boxplots for both methods under H_0 and H_1. In scenario 3, $n_1 = 140$, $n_2 = 130$, $n_3 = 120$, $n_4 = 40$, and $n_5 = 60$ are the sample sizes. Figure A.13 shows the RP value and boxplots for both hypotheses under H_0 and H_1. In both cases of rejection and non-rejection, a boxplot is presented. The RP increases when the p-value is moved away from the test threshold. Clearly, reproducibility is lowest around the threshold of the test. It has been shown that Bayes-RP and NPI-B-RP become closer to 0.5 with increasing sample size when the observed p-value is near to the test threshold in both rejection and non-rejection cases. Increasing the sample size leads to these results since the variability of both mothed samples decreases and the power of the test increases when the sample size is increased. In general, as sample sizes increase, both rejection and non-rejection probabilities tend to converge towards 0.5 when p-values fall close to test threshold, and variability in RP values decreases, primarily due to the more powerful test that results from a larger sample.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

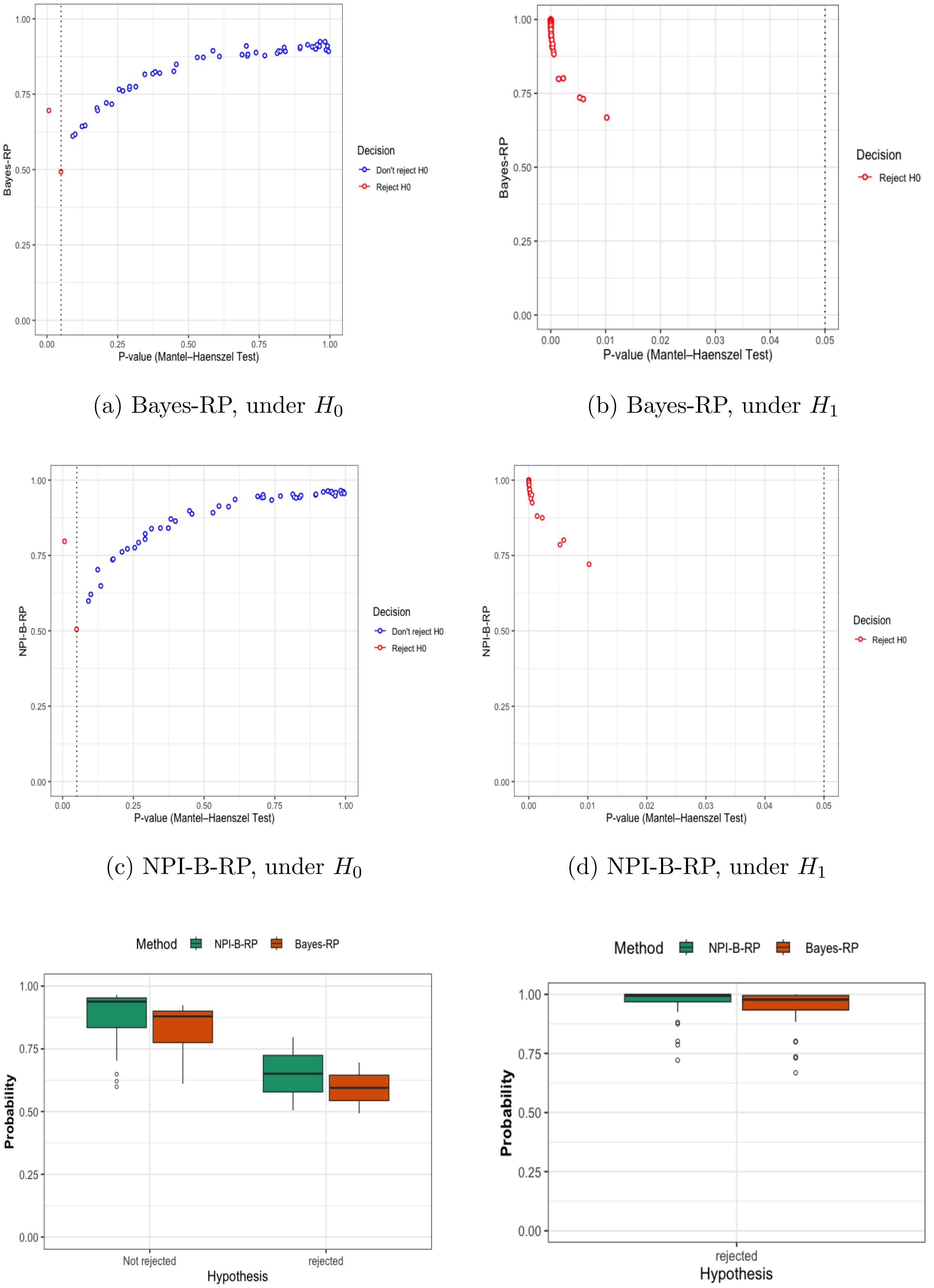
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure A.11: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Mantel-Haenszel test , Scenario 1 for $k = 5$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

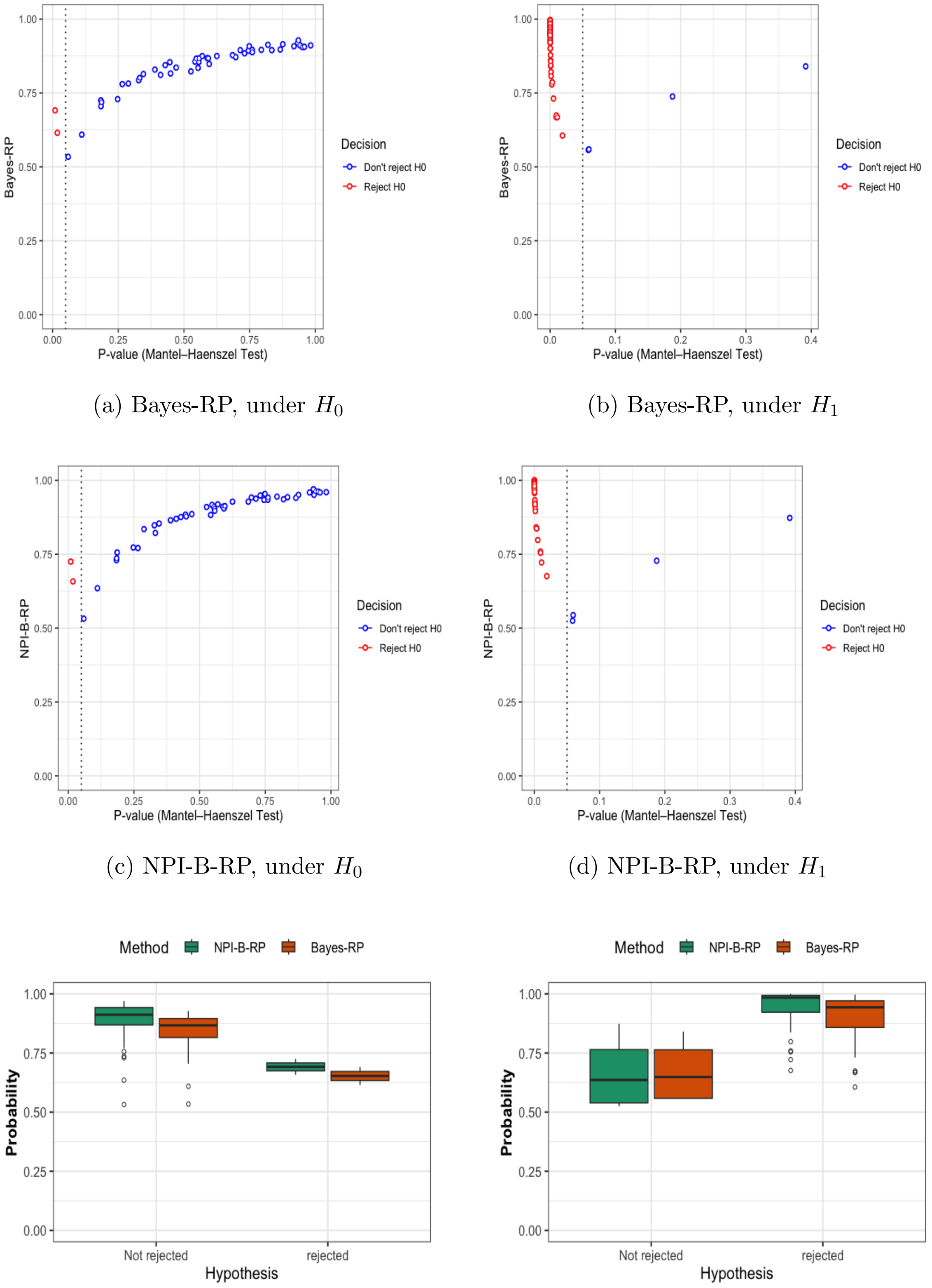
(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure A.12: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP values for Mantel-Haenszel test , Scenario 2 for $k = 5$.



(a) Bayes-RP, under H_0

(b) Bayes-RP, under H_1

(c) NPI-B-RP, under H_0

(d) NPI-B-RP, under H_1

(e) RP, under H_0

(f) RP, under H_1

Figure A.13: Simulations under H_0 and H_1: Bayes-RP and NPI-B-RP valuesfor Mantel-Haenszel test , Scenario 3 for $k = 5$.

References

[1] Agresti A. (2002). *Categorical Data Analysis.* Wiley, New York.

[2] Agresti A. (2007). *Categorical Data Analysis.* Wiley, New York.

[3] Agresti A. and Coull B.A. (1998). Approximate is better than “exact” for interval estimation of binomial proportions. *The American Statistician*, 52(2), 119–126.

[4] Albert A. (2017). Biostatistics: Facing the Interpretation of 2× 2 Tables. *Journal of the Belgian Society of Radiology*, 101(Suppl 2).

[5] Albert J.H. and Gupta A.K. (1983). Bayesian estimation methods for 2× 2 contingency tables using mixtures of Dirichlet distributions. *Journal of the American Statistical Association*, 78(383), 708–717.

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

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

[8] Augustin T., Coolen F.P.A., De Cooman G. and Troffaes M.C. (2014). *Introduction to Imprecise Probabilities.* John Wiley & Sons Hoboken, New Jersey.

[9] Begley C. and Ellis L. (2012). Drug development: Raise standards for preclinical cancer research. Nature. 483 (7391).

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

[11] BinHimd S. (2014). Nonparametric predictive methods for bootstrap and test reproducibility. *PhD thesis*, Durham University, https://etheses.dur.ac.uk.

[12] Bodden C., von Kortzfleisch V.T., Karwinkel F., Kaiser S., Sachser N. and Richter S.H. (2019). Heterogenising study samples across testing time improves reproducibility of behavioural data. *Scientific Reports*, 9(1), 8247.

[13] Bolstad W.M. and Curran J.M. (2016). *Introduction to Bayesian Statistics.* John Wiley & Sons, New York.

[14] Breslow N.E., Day N.E. and Heseltine E. (1980). Statistical methods in cancer research.

[15] Carlin J. and Doyle L. (2001). Comparison of means and proportions using confidence intervals. *Journal of Paediatrics and Child Health*, 37(6), 583–586.

[16] Chow S.C., Shao J., Wang H. and Lokhnygina Y. (2017). *Sample Size Calculations in Clinical Research.* Chapman and Hall/CRC, Boca Raton, Florida.

[17] Cochran W.G. (1954). Some methods for strengthening the common χ 2 tests. *Biometrics*, 10(4), 417–451.

[18] Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences.* Lawrence Erlbaum Associates, Hillsdale, NJ.

[19] Coolen F.P.A. (1998). Low structure imprecise predictive inference for Bayes' problem. *Statistics & Probability Letters*, 36(4), 349–357.

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

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

[22] 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(2), 217–230.

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

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

[25] Coolen F.P.A., Troffaes M.C. and Augustin T. (2011). Imprecise Probability. *International Encyclopedia of Statistical Science*, pp. 645–648.

[26] Coolen F.P.A. and Yan K. (2004). Nonparametric predictive inference with right-censored data. *Journal of Statistical Planning and Inference*, 126(1), 25–54.

[27] Cremers J. and Klugkist I. (2018). One direction? A tutorial for circular data analysis using R with examples in cognitive psychology. *Frontiers in Psychology*, 9, 2040.

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

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

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

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

[32] Di Serio C., Malgaroli A., Ferrari P. and Kenett R.S. (2022). The reproducibility of COVID-19 data analysis: paradoxes, pitfalls, and future challenges. *PNAS Nexus*, 1(3).

[33] Efron B. and Tibshirani R.J. (1993). An introduction to the bootstrap Chapman & Hall. *New York*, 436.

[34] Etz A. and Vandekerckhove J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *Plos One*, 11(2), e0149794.

[35] Fagerland M., Lydersen S. and Laake P. (2017). *Statistical Analysis of Contingency Tables.* Chapman and Hall/CRC, New York.

[36] Gelman A., Carlin J.B., Stern H.S. and Rubin D.B. (1995). *Bayesian Data Analysis.* Chapman and Hall/CRC, Boca Raton, FL.

[37] Gelman A., Carlin J.B., Stern H.S. and Rubin D.B. (2014). Bayesian data analysis (Vol. 2).

[38] Good I.J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4), 237–264.

[39] Good I.J., Hacking I., Jeffrey R. and Törnebohm H. (1966). The estimation of probabilities: An essay on modern Bayesian methods. *Synthese*, 16(2).

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

[41] Hill B.M. (1968). Posterior distribution of percentiles: Bayes' theorem for sampling from a population. *Journal of the American Statistical Association*, 63(322), 677–691.

[42] Hill B.M. (1988). De Finetti's theorem, induction, and A (n) or Bayesian nonparametric predictive inference (with discussion). *Bayesian statistics*, 3, 211–241.

[43] Hoadley B. (1969). The compound multinomial distribution and Bayesian analysis of categorical data from finite populations. *Journal of the American Statistical Association*, 64(325), 216–229.

[44] Hoff P.D. (2009). *A First Course in Bayesian Statistical Methods*, volume 580. Springer, New York.

[45] Höpfl S., Pleiss J. and Radde N.E. (2023). Bayesian estimation reveals that reproducible models in Systems Biology get more citations. *Scientific Reports*, 13(1), 2695.

[46] Howard J. (1998). The 2× 2 table: A discussion from a Bayesian viewpoint. *Statistical Science*, pp. 351–367.

[47] Jarvis M.F. and Williams M. (2016). Irreproducibility in preclinical biomedical research: perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences*, 37(4), 290–302.

[48] Kateri M. (2014). *Contingency Table Analysis: Methods and Implementation Using R.* Springer, New York.

[49] Kateri M. (2018). ϕ-divergence in contingency table analysis. *Entropy*, 20(5), 324.

[50] Kateri M. and Agresti A. (2013). Bayesian inference about odds ratio structure in ordinal contingency tables. *Environmetrics*, 24(5), 281–288.

[51] Kirkwood B.R. and Sterne J.A. (2010). *Essential Medical Statistics.* John Wiley & Sons, New York.

[52] Kruschke J.K. (2021). Bayesian analysis reporting guidelines. *Nature Human Behaviour*, 5(10), 1282–1291.

[53] Lecoutre B. and Charron C. (2000). Bayesian procedures for prediction analysis of implication hypotheses in 2× 2 contingency tables. *Journal of Educational and Behavioral Statistics*, pp. 185–201.

[54] Lindley D.V. (1972). *Bayesian Statistics: A review.* Society for industrial and applied mathematics.

[55] Mantel N. and Haenszel W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719–748.

[56] 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, 1–14.

[57] McNemar Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2), 153–157.

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

[59] Murphy K.P. (2012). *Machine Learning: a Probabilistic Perspective.* MIT press, Cambridge, MA.

[60] Paez A. (2022). Reproducibility of research during COVID-19: Examining the case of population density and the basic reproductive rate from the perspective of spatial analysis. *Geographical Analysis*, 54(4), 860–880.

[61] Peers I.S., South M.C., Ceuppens P.R., Bright J.D. and Pilling E. (2014). Can you trust your animal study data? *Nature Reviews Drug Discovery*, 13(7), 560–560.

[62] Plant A. and Hanisch R. (2020). Reproducibility in science: A metrology perspective.

[63] Sauro J. and Lewis J.R. (2005). Estimating completion rates from small samples using binomial confidence intervals: comparisons and recommendations. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 49, pp. 2100–2103. SAGE Publications Sage CA: Los Angeles, CA.

[64] Schwalbe M. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop.* National Academies Press.

[65] Senn S. (2002). A comment on replication, p-values and evidence SN Goodman. *Statistics in Medicine*, 21(16), 2437–2444.

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

[67] Simkus A., Coolen F.P.A., Coolen-Maturi T., Karp N.A. and Bendtsen C. (2022). Statistical reproducibility for pairwise t-tests in pharmaceutical research. *Statistical Methods in Medical Research*, 31(4), 673–688.

[68] Stahel W.A. (2021). New relevance and significance measures to replace p-values. *PLoS One*, 16(6), e0252991.

[69] Stanley T.D., Carter E.C. and Doucouliagos H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin*, 144(12), 1325.

[70] Stodden V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application*, 2, 1–19.

[71] Stone J.V. (2013). Bayes' rule: a tutorial introduction to Bayesian analysis.

[72] Sy K.T.L., White L.F. and Nichols B.E. (2023). Reproducible Science Is Vital for a Stronger Evidence Base During the COVID-19 Pandemic. *Geographical Analysis*, 55(1), 203–206.

[73] Walley P. (1991). *Statistical Reasoning with Imprecise Probabilities*, volume 42. Springer, New York.

[74] Witte R.S. and Witte J.S. (2017). *Statistics.* John Wiley & Sons, New York.

[75] Woolf B. *et al.* (1955). On estimating the relation between blood group and disease. *Ann hum genet*, 19(4), 251–253.

[76] Zelen M. (1971). The analysis of several 2× 2 contingency tables. *Biometrika*, 58(1), 129–137.