Reproducibility of Statistical Inference Based on Randomised Response Data

ALGHAMDI, FATIMAH,MOHAMMAD

Reproducibility of Statistical Inference Based on Randomised Response Data

Fatimah M. Alghamdi

A Thesis presented for the degree of
Doctor of Philosophy

The Statistics Group
Department of Mathematical Sciences
University of Durham
England

October 2022

Dedicated to

My lovely children: Meshari, Muhannad, and Mira.

Reproducibility of Statistical Inference Based on Randomised Response Data

Fatimah M. Alghamdi

Submitted for the degree of Doctor of Philosophy
October 2022

Abstract

Reproducibility of an experiment's conclusion is an important topic in a variety of fields, including social studies. This thesis presents a theory of reproducibility of statistical inference based on randomised response data. First, reproducibility of statistical hypothesis tests based on randomised response data is studied. This thesis presents statistical inference for reproducibility of the outcome of a hypothesis test based on data resulting from different randomised response techniques (RRT). Secondly, a new method for quantifying reproducibility of statistical estimates is introduced. Finally, this method is applied to derive reproducibility of estimates of population characteristics based on randomised response data.

The quantification of reproducibility uses nonparametric predictive inference (NPI), which is suitable for reproducibility when considering this as a prediction problem. NPI uses only few model assumptions and results in lower and upper reproducibility probabilities. We compared different randomised response methods. The results of this thesis open up the possibility of pre-selecting a randomised response method with higher reproducibility and also indicate the relationship between variance and reproducibility with the same privacy level. We find that less variability in the reported responses of RRT methods leads to higher reproducibility of statistical hypothesis tests based on RRT data with the same privacy degree.

Therefore, for RRT methods using binary responses, reproducibility of hypothesis tests based on the forced method is greater than reproducibility of hypothesis tests based on the Greenberg method. For RRT methods using real-valued responses, reproducibility of estimates is greater for data collected from the Greenberg method than the reproducibility of estimates for data collected from the optional multiplicative method and the Eichhorn and Hayre method.

Declaration

The work in this thesis is based on research carried out by the Department of Mathematical Sciences at Durham University, UK. 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

For the numerous blessings, I am very grateful to Allah, my God. He has manifested himself to me in a number of ways, including in the completion of this thesis.

I would like to express my heartfelt gratitude to Prof. Frank Coolen and Dr Tahani Coolen-Maturi, my academic supervisors, who deserve my heartfelt gratitude and deepest admiration for their unwavering support, patience, compassion, excitement, calm advise, recommendations, and direction throughout my work and during the preparation of my thesis. They are really pleasant and friendly individuals who have greatly assisted and encouraged me during my time in the UK. All of my phrases are not entirely enough; however, they do provide excellent support, particularly in terms of prompt responses and letter writing, even on their weekends or holiday. No matter what I say, I will not be able to give them right praise, thanks and gratitude.

My father and mother have been a huge source of inspiration, continuous support, adoration, and supplications, and have consistently stood by me and been pleased with me, and I am very grateful to them. I would want to express my heartfelt gratitude to my brothers and sisters for their unwavering faith and support. My inspired brother Ahmed Alghamdi deserves special thanks for his wise counsel and unwavering support during tough moments in my life.

My husband deserves special thanks and gratitude for his perseverance, adoration, support, standing by me, and generosity. Meshari, Muhannad, and Mira, please accept my heartfelt gratitude.

Thank Prof. Fatimah Al-Aboudi very much for her encouragement. Your words stayed

with me. Those remarks, which were a ray of optimism in the midst of my PhD studies will stay with me for the rest of my life.

I am grateful for the financial and psychological support provided by Princess Nourah bint Abdulrahman University in Saudi Arabia, as well as the Saudi Arabian Cultural Bureau and all of their London staff, which enabled me to pursue my PhD studies at Durham University.

Thank you very much to Durham University and all their Staff for providing me with a welcoming academic environment and for providing me with amazing opportunities to pursue my interests and present my findings at international conferences. Finally, I owe a debt of gratitude to my colleagues, companions, and everyone else who has helped and supported me or contributed to my progress.

Contents

Chapter 1

Introduction

1.1 Overview

In social research, it is often necessary to ask questions on issues that could be sensitive to the respondent, in order to obtain truthful answers from the people to whom the research study is applied. Randomised Response Technique (RRT) is an effective method which helps to elicit the truth. It is critical with this technique to obtain accurate responses to sensitive questions from respondents while maintaining their privacy. There are various distinct RRT approaches, some qualitative and others quantitative; the Warner method [95] is the first method that was introduced as explained in the literature review in Section 1.2.

Social science statistical inference allows for the estimation from a sample of the population that is impacted by a certain phenomenon. Because people differ over time, between geographies, and in relation to social life, we could reach at a different conclusion if we repeat the experiment. By being observant of how much the outcomes could change if we repeated the test, we can take this difference into account when drawing inferences. Additionally, it allows for the statement of whether or not this social study provides evidence to reject a hypothesis and to provide a reasonable range for the true value of any property in the population, such as the population's proportion of this property.

This thesis presents different methods to investigate reproducibility of statistical inference. We focus on three main topics: reproducibility, nonparametric predictive inference (NPI), and randomised response methods (RRT).

The first contribution, we investigate the reproducibility of statistical hypothesis tests based on RRT data using the one-sided and the two-sided test. The reproducibility of statistical hypothesis tests has received attention according to De Capitani and De Martini [44] who were the first researchers that investigated reproducibility estimators for a number of nonparametric tests, including the sign, and Wilcoxon signed rank tests.

Reproducibility is concerned with the question of whether a second statistical test performed under identical conditions will provide the same result as the original tests in terms of rejection or non-rejection of the null hypothesis.

We study the reproducibility of statistical hypothesis tests using Nonparametric Predictive Inference (NPI). The NPI method is a frequentist statistics framework that focus on future observations that are exchangeable with actual observations. NPI is based on only few modelling assumptions. NPI has ability to predict which makes it suitable method for determining test reproducibility.

The second contribution is reproducibility of estimates. Of course, the original estimates of distribution characteristics for real-valued random quantities will differ from the future estimates, which means that the future estimates will not lead to the exact same value as the original estimates. Therefore, we consider reproducibility for estimates in terms of the difference between actual estimates and estimates based on a future data set.

The third contribution is the application of reproducibility for estimates with data collected using randomised response methods.

NPI for reproducibility of statistical tests based on RRT data (NPI-RP-RRT) and for reproducibility of estimates will be presented using several RRT in order to compare them in terms of reproducibility, while we also consider the efficiency and privacy of these RRT methods.

The primary ideas and concepts employed in this thesis are explained in this chapter as follows. Section 1.2 provides a literature review of a variety of RRT methods. Section 1.3, explains the reproducibility concept. Section 1.4 introduces nonparametric predictive inference, NPI for Bernoulli random quantities and NPI-B method. Then, the NPI for reproducibility is introduced in Section 1.5.

1.2 Randomised response techniques (RRT)

Randomised response techniques (RRT) are methods to elicit truth responses from respondents to sensitive questions in a survey. To avoid embarrassment when respondents are asked sensitive questions. A spinner, a deck of cards, or a coin can be used as randomisation device and the responses are hidden from the interviewer. These methods conceal individual responses and maintains respondent privacy by generating randomness.

There are several forms of RRT available to eliminate bias caused by respondents' hesitancy or respondents providing incorrect responses, which affects the accuracy of the results. Warner [95] presented the first RRT method which we refer to as the Warner method (WM).

The WM method is illustrated as follows. Suppose that we have a population and we wish to estimate the proportion π of people who have a sensitive characteristic A. We have two questions, the sensitive question Q_1 and the non-sensitive question Q_2, to determine if the respondent is in the target group A (they have the sensitive characteristic) or if they do not have the sensitive characteristic $\bar{A}$ as follows.

$$Q_1 : \quad \text{Are you a member of group A?}$$
$$Q_2 : \quad \text{Are you not a member of group A?}$$

Let there be a randomisation device to help respondents to choose the question and then answer it. Suppose that with probability γ, we have the sensitive questions and with probability $1-\gamma$, we have the non-sensitive question, where γ is only known of the interviewer. As a result, the number of people who get the sensitive question is binomially distributed with sample size n and parameter γ. Each response can result in one of two possible outcomes, a Yes-answer ($\dot{Y}$) or a No-answer ($\dot{N}$) regardless of the question selected.

Assume that Y is the binomial random quantity of the number of people who will answer 'Yes' to the sensitive question where possible answers are only 'yes' or 'no'. If the probability of a 'yes' answer is given by:

$$P^* = \gamma\pi + (1-\gamma)(1-\pi) \tag{1.1}$$

then, the expected value of Y is $\mathrm{E}(Y) = nP^*$.

The estimator of the proportion $\hat{\pi}(Y)$ of people who have the sensitive characteristic is

$$\hat{\pi}(Y) = \frac{n(\gamma-1)+Y}{(2\gamma-1)n} \quad \text{where} \quad 0 \leq \gamma \leq 1 \;\; \text{and} \;\; \gamma \neq \frac{1}{2} \tag{1.2}$$

The expectation of the estimator $\hat{\pi}(Y)$ is

$$\mathrm{E}(\hat{\pi}(Y)) = \mathrm{E}\left[\frac{n(\gamma-1)+Y}{(2\gamma-1)n}\right] = \pi \tag{1.3}$$

So $\hat{\pi}(Y)$ is an unbiased estimator of π. The variance of the estimator $\hat{\pi}(Y)$ is [95]:

$$\mathrm{Var}(\hat{\pi}(Y)) = \frac{(\pi-\pi^2)}{n} + \frac{\gamma(1-\gamma)}{n(2\gamma-1)^2} \quad \text{where} \quad 0 \leq \gamma \leq 1,\; \gamma \neq \frac{1}{2} \tag{1.4}$$

The first term in Equation (1.4) is the binomial variance related to a sensitive question. The second term is the extra variance for the uncertainty caused by using the randomisation device, which becomes substantial if γ is close to 0.5.

In this method, Warner [95] suggested that the probability of a sensitive question in the randomisation device should be greater than 0.5 which is the point of interest in this method. The reason for this choice is that if $\gamma = 0.5$, then the probability of person i who says 'Yes' will not depend on π in Equation (1.1), then the data would hold no information about π. If $\gamma = 1$, we just return to the non-RRT methods and use the direct question. If we choose $0.5 < \gamma \leq 1$, the respondent provides a useful response and the respondent never refers to which group they belong. Therefore, γ can describe whether the respondent cooperates, by answering any question asked, or not.

As a result, good selections of γ and n are essential to provide a level of accuracy of P^* and standard deviation of the estimator $\hat{\pi}(Y)$. For example, if $\pi = 0.5$ and $\gamma = 0.75$, the variance given by Equation (1.4) equals $1/n$. However, to achieve an accurate result of P^*, the sample size should be 400 to ensure that the standard deviation is 0.05 which indicates that under the regular method, each response was truthful. Comparatively, the classical estimating approach (corresponding to $\gamma = 1$) would suggest that only a sample of about 100 would be required for a standard deviation of 0.05 [95].

More widely, it should be noted that there appear to be large potential improvements from the randomised response, excepting situations where the bias of the regular estimate is minimal or 0. For example, using larger samples such as 2000 leads to reducing the mean square error of the estimates. In addition, using a γ as low as 0.6 is needed to ensure collaboration from the respondents [95].

After creating the WM, several researchers presented a variety of randomised response techniques to reduce bias and obtain accurate responses from respondents, with some focusing on the type of questions used in the process and others focusing on the usage of various shapes of randomisation devices and others focus on the type of the responses such as the binary and real-valued responses of RRT as explained in Sections 1.2.1 and 1.2.2.

1.2.1 Qualitative randomised response techniques

In this section, we introduce qualitative RRT for surveys in which sensitive questions are answered using qualitative binary response variables such as ‘Yes’ or ‘No’. The Greenberg technique and the forced approach are two RRT approaches that use binary responses and are substantially used in this thesis.

The Greenberg Method (GB) [60] is a variation of the WM in which respondents are also randomly assigned to one of two questions using the randomisation device. With known probability γ, the respondent is asked the question about the sensitive issue, and with probability $1 - \gamma$, the respondents are asked an unrelated question and not sensitive.

Assume that we have a sample of size n, and random quantity Y as the number of people who answer 'Yes'. Let π_A be the proportion of people who have the sensitive characteristic, and π_B is the proportion of people who would respond 'Yes' to the unrelated question. It is assumed that π_B is known. Because the characteristic B is not a sensitive feature, therefore, we assume that respondents answer question B truthfully. The two questions could be:

$$\begin{aligned} Q_1 :& \quad \text{Are you a member of group A?} \\ Q_2 :& \quad \text{Are you a member of group B?} \end{aligned}$$

Then, the probability of the event that a person answers 'Yes' to the question

$$P^* = \gamma \pi_A + (1-\gamma)\pi_B \tag{1.5}$$

Note that, as for WM, in applying GB, the interviewer is unaware of the question being asked.

It is preferable to choose the unrelated characteristic B with probability π_B that is not close to zero. Such action could contradict the core purpose of using the unrelated question approach because choosing π_B close to zero could affect the respondent's desire to respond truthfully, a good rule is to aim for π_B in the neighbourhood of 0.10. If π_A is very small, say 0.01, it is not always desirable to choose π_B in the neighbourhood of 0.10, even if such a choice is advantageous theoretically based on the sampling variance [2].

Let Y be the random quantity of the number of people in the sample of size n who answer 'Yes' to the two questions they are asked, then the estimator of proportion $\hat{\pi}_A(Y)$ of people who have the sensitive characteristic is

$$\hat{\pi}_A(Y) = \frac{\frac{Y}{n} - \pi_B(1-\gamma)}{\gamma} \tag{1.6}$$

The expected value of the estimator $\hat{\pi}_A(Y)$ is

$$\begin{aligned} \mathrm{E}(\hat{\pi}_A(Y)) &= \mathrm{E}\left(\frac{\frac{Y}{n} - (1-\gamma)\pi_B}{\gamma} \right) = \frac{P^* - (1-\gamma)\pi_B}{\gamma} \\ &= \frac{\gamma\pi_A + (1-\gamma)\pi_B - (1-\gamma)\pi_B}{\gamma} = \pi_A \end{aligned}$$

where $\hat{\pi}_A$(Y) is an unbiased estimate of the population proportion π_A.

The variance of $\hat{\pi}_A(Y)$ is [100]:

$$\begin{aligned}
\mathrm{Var}(\hat{\pi_A}(Y)) &= \mathrm{Var}\left(\frac{P^* - (1-\gamma)\pi_B}{\gamma}\right) = \mathrm{Var}(\frac{P^*}{\gamma}) \\
&= \frac{[\pi_A\gamma + \pi_B(1-\gamma)][1-(\pi_A\gamma + \pi_B(1-\gamma))]}{n\gamma^2} \\
&= \frac{-\pi_A^2\gamma^2 + 2\pi_A\pi_B\gamma^2 - \pi_B^2\gamma^2 - 2\pi_A\pi_B\gamma}{n\gamma^2 + 2\pi_B\gamma} + \frac{\pi_A\gamma - \pi_B^2 - \pi_B\gamma + \pi_B}{n\gamma^2} \\
&= \frac{-\pi_A^2\gamma^2 + 2\pi_A\pi_B\gamma^2 - \pi_B^2\gamma^2 - 2\pi_A\pi_B\gamma}{n\gamma^2 + 2\pi_B\gamma} \\
&+ \frac{\pi_A\gamma - \pi_B^2 - \pi_B\gamma + \pi_B}{n\gamma^2} + \frac{\pi_B\gamma^2 - \pi_B\gamma^2 + \pi_A\gamma^2 - \pi_A\gamma^2}{n\gamma^2} \\
&= \frac{\pi_A(1-\pi_A)}{n} + \frac{(1-\gamma)^2\pi_B(1-\pi_B) + \gamma(1-\gamma)(\pi_A + \pi_B - 2\pi_A\pi_B)}{n\gamma^2} \qquad (1.7)
\end{aligned}$$

where $0 \leq \gamma \leq 1$ and $\gamma \neq \frac{1}{2}$, and the term in the first equality $\mathrm{Var}\left(\frac{(1-\gamma)\pi_B}{\gamma}\right)$ is equal 0 because it is constant.

The Forced Method (FM) [19] is a simple implementation of an RRT. When using the forced response method, the randomisation device forces the respondent to answer 'Yes' with probability γ_1 or 'No' with probability γ_2 or to answer the sensitive question with probability γ truthfully, where $\gamma = 1 - \gamma_1 - \gamma_2$ and $0 < \gamma_1 < 1$, $0 < \gamma_2 < 1$ and $\gamma_1 + \gamma_2 < 1$ [19].

Assume that we have a sample of size n, and random quantity Y denoting the number of people who answer 'Yes'. The probability of a respondent answering 'Yes' using the sensitive question or forced Yes-response is

$$P^* = \gamma_1 + \pi_A(1 - \gamma_1 - \gamma_2) \qquad (1.8)$$

where π_A again the proportion of people who have the sensitive characteristic A. The estimator of the proportion of people who have the sensitive characteristic is

$$\hat{\pi}_A(Y) = \frac{\frac{Y}{n} - \gamma_1}{1 - \gamma_1 - \gamma_2} \qquad (1.9)$$

and the expected value of $\hat{\pi}_A(Y)$ is

$$\begin{aligned}
\mathrm{E}(\hat{\pi}_A(Y)) &= \mathrm{E}\left(\frac{\frac{Y}{n} - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) \\
&= \mathrm{E}\left(\frac{P^* - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) \\
&= \mathrm{E}\left(\frac{\gamma_1 + \pi_A(1 - \gamma_1 - \gamma_2) - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) \\
&= \pi_A \qquad (1.10)
\end{aligned}$$

where $\hat{\pi}_A$(Y) is a unbiased estimate of the population proportion π_A. The variance of the estimator $\hat{\pi}_A(Y)$ is [19]:

$$\begin{aligned}
\mathrm{E}(\hat{\pi}_A(Y)) &= \mathrm{E}\left(\frac{\frac{Y}{n} - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) \\
&= \mathrm{E}\left(\frac{P^* - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) \\
&= \mathrm{E}\left(\frac{\gamma_1 + \pi_A(1 - \gamma_1 - \gamma_2) - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) \\
&= \pi_A \qquad (1.11)
\end{aligned}$$

where $\hat{\pi}_A$(Y) is a unbiased estimate of the population proportion π_A. The variance of the estimator $\hat{\pi}_A(Y)$ is [19]:

$$\begin{aligned}
\mathrm{Var}(\hat{\pi}_A(Y)) =& \mathrm{Var}\left(\frac{\frac{Y}{n} - \gamma_1}{1 - \gamma_1 - \gamma_2}\right) = \mathrm{Var}\left(\frac{P^*}{n(1 - \gamma_1 - \gamma_2)^2}\right) \\
=& \frac{[\gamma_1 + \pi_A(1 - \gamma_1 - \gamma_2)][1 - (\gamma_1 + \pi_A(1 - \gamma_1 - \gamma_2))]}{n(1 - \gamma_1 - \gamma_2)^2} \\
=& \frac{\gamma_1 - \gamma_1^2 - \pi_A\gamma_1(1 - \gamma_1 - \gamma_2) + \pi_A(1 - \gamma_1 - \gamma_2)}{n(1 - \gamma_1 - \gamma_2)^2} \\
-& \frac{\pi_A\gamma_1(1 - \gamma_1 - \gamma_2) - \pi_A^2(1 - \gamma_1 - \gamma_2)^2}{n(1 - \gamma_1 - \gamma_2)^2} \\
=& \frac{\pi_A[-\gamma_1(1 - \gamma_1 - \gamma_2) + (1 - \gamma_1 - \gamma_2) - \gamma_1(1 - \gamma_1 - \gamma_2)]}{n(1 - \gamma_1 - \gamma_2)^2} \\
-& \frac{\pi_A^2(1 - \gamma_1 - \gamma_2)^2 - \gamma_1(1 - \gamma_1)}{n(1 - \gamma_1 - \gamma_2)^2} \\
=& \frac{\pi_A[(1 - 2\gamma_1 - \gamma_2 + \gamma_2)(1 - \gamma_1 - \gamma_2)]}{n(1 - \gamma_1 - \gamma_2)^2} - \frac{\pi_A^2(1 - \gamma_1 - \gamma_2)^2 - \gamma_1(1 - \gamma_1)}{n(1 - \gamma_1 - \gamma_2)^2} \\
=& \frac{\pi_A(\pi_A - 1)}{n} + \frac{\pi_A(\gamma_2 - \gamma_1)}{n(1 - \gamma_1 - \gamma_2)} + \frac{\gamma_1(1 - \gamma_1)}{n(1 - \gamma_1 - \gamma_2)^2} \qquad (1.12)
\end{aligned}$$

Other RRT methods have been proposed, each with specific procedures and assumptions, such as scenarios in which respondents are truthful or untruthful in their responses, or using multiple randomisation devices. For more details, we refer to [4,19,53,54,75,80,84,85].

1.2.2 Quantitative randomised response techniques

This section introduces some randomised response techniques for quantitative responses which uses real numbers to response the questions such as the Greenberg method, the Eichhorn and Hayre method, the optional multiplicative method and the additive method.

The Greenberg method (GM) [61] is a quantitative variation of the unrelated question approach (GB) for quantitative responses. Respondents utilise the randomisation device to answer one of two questions. One of these questions is sensitive while the other is nonsensitive. Both answers of these question are real-valued quantities.

Assume the probability of the sensitive question is γ, and denote answer as random quantity X_i with expected value $E(X_i) = \mu_x$ and variance σ_x^2. The probability of the unrelated question is $1-\gamma$, and this is an unrelated non-sensitive question. Let denote the answer as random quantity Y_i with expected value $E(Y_i) = \mu_y$ and variance σ_y^2. Both μ_y and σ_y^2 are assumed to be known, because if it is not we need to estimate them.

Let assume that random quantity Z_i denotes response of the i^{th} respondent ($i = 1, 2, ..., n$), so

$$Z_i = \begin{cases} X_i & \text{with probability} \quad \gamma \\ Y_i & \text{with probability} \quad 1-\gamma \end{cases} \tag{1.13}$$

Then, the expected value of Z_i is

$$\begin{aligned} E(Z_i) =& \gamma E(X_i) + (1-\gamma)E(Y_i) \\ =& \gamma\mu_x + (1-\gamma)\mu_y \end{aligned} \tag{1.14}$$

with $\bar{Z}$ regarding to the sample mean, we can estimate μ_x by

$$\hat{\mu}_x = \frac{\bar{Z} - (1-\gamma)\mu_y}{\gamma} \tag{1.15}$$

The variance of Z_i [98] is

$$\begin{aligned}
\mathrm{Var}(Z_i) &= [\gamma E(Z_i^2) + (1-\gamma)E(Z_i^2)] - (E(Z_i))^2 \\
&= \frac{1}{\gamma^2}\left[\gamma(\mu_x^2 + \sigma_x^2) + (1-\gamma)(\sigma_y^2 + \mu_y^2) - \{\gamma\mu_x + (1-\gamma)\mu_y\}^2\right] \\
&= \frac{1}{\gamma^2}\left[\sigma_y^2 + \gamma(\sigma_x^2 - \sigma_y^2) + \gamma(1-\gamma)\mu_x^2 - \gamma(1-\gamma)\mu_y^2 - 2\gamma(1-\gamma)\mu_x\mu_y\right] \\
&= \frac{1}{\gamma^2}\left[\sigma_y^2 + \gamma(\sigma_x^2 - \sigma_y^2) + \gamma(1-\gamma)(\mu_x - \mu_y)^2\right]
\end{aligned}$$

Eichhorn and Hayre method (EH) is a quantitative approach which relies on scrambled responses rather than true responses and can be implemented by adding, removing, or multiplying real responses by random numbers.

Eichhorn and Hayre [51] presented a full randomisation multiplicative scrambling approach in which respondents were asked to product their response with a scrambling number using the randomisation device.

Assume that we have random quantity X_i as the true response with expected value $E(X_i) = \mu_x$ and variance $\sigma_x^2 = V(X_i)$, where $i = 1, ..., n$, and μ_x and σ_x^2 are unknown. The randomisation device provides a numerical value S_i that follows a predetermined probability distribution with a known mean $E(S_i) = \theta$ and variance r^2, where the random quantities S_i and X_i are assumed to be independent variables. In this method, respondents choose a number and report the product of the real responses X_i and S_i, as follows:

$$Z_i = X_i S_i \tag{1.16}$$

Because S_i and X_i are assumed to be independent, we have:

$$E(Z_i) = E(X_i)E(S_i) = \mu_x\theta \tag{1.17}$$

The variance of the unbiased estimator of the sensitive characteristic's mean $\hat{\mu}_x$ is

$$\begin{aligned}\text{Var}(\hat{\mu_x}) =& \frac{1}{n}\text{Var}(\frac{\bar{Z}}{\theta}) = \frac{1}{n\theta^2}\left[E(Z_i^2) - E(Z_i)^2\right] \qquad (1.18)\\ =& \frac{1}{n\theta^2}\left[E(S_iX_i)^2 - E(X_i)^2E(S_i)^2\right]\\ =& \frac{1}{n\theta^2}\left[E[S_i^2]E[X_i^2] - \mu_x^2\theta^2\right]\\ =& \frac{1}{n\theta^2}\left[(r^2+\theta^2)(\sigma_x^2+\mu_x^2) - \mu_x^2\theta^2\right]\\ =& \frac{1}{n}\left[\sigma_x^2 + \frac{r^2}{\theta^2}(\sigma_x^2+\mu_x^2)\right] \qquad (1.19)\end{aligned}$$

where $\hat{\mu_x} = \frac{\bar{Z}}{\theta}$ and $\bar{Z} = \sum_{i=1}^{n}\frac{Z_i}{n}$.

The optional multiplicative method (MM) is another quantitative method. Gupta [62] developed multiplicative optional scrambling of RRT method, in which an unknown proportion of respondents scramble their responses as sensitive, other respondents do not consider the issue sensitive and give their true responses. When adopting this approach, respondents are not required to scramble their responses if they do not think the issue is sensitive.

Assume that we have random quantity X_i as a sensitive characteristic for individual i with an unknown mean μ_x, and random quantity S_i as a scrambling variable with a known mean $E(S_i)$, where X_i and S_i are independent, and S_i can be produced from any distribution. Let's assume that we have a random quantity Z_i denoting the response of a person i where $i = 1, ..., n$. Giving the randomisation device which gives a random quantity S_i that follows a known probability distribution with the known mean $E(S_i) = 1$ and known variance γ^2. S_i and X_i are both random variables with positive values, therefore we assume that $E(S_i) = 1$.

The respondent offers the answer $Z_i = X_i$ if the question is not sensitive; if the question is sensitive, the answer is scrambling $Z_i = S_iX_i$. Each respondent has an equal probability of being chosen. All respondents have the same chance of scrambling ψ, which is a known quantity. Therefore, the reported responses Z_i is

$$Z_i = \begin{cases} X_i & \text{with probability} \quad \psi \\ X_i S_i & \text{with probability} \quad 1-\psi \end{cases} \tag{1.20}$$

Under the assumption that X_i and S_i are independent, the expected value of Z_i is as follows.

$$\begin{aligned} E(Z_i) = \mu_z =& \psi E(X_i) + (1-\psi)E(X_i)E(S_i) \\ =& \psi E(X_i) + E(X_i)E(S_i) - \psi E(X_i)E(S_i) \\ =& \psi E(X_i) + E(X_i) - \psi E(X_i) \\ =& E(X_i) = \mu_x \end{aligned} \tag{1.21}$$

where $E(S_i) = 1$, then $\mu_z = \mu_x$. That means the estimator $\hat{\mu}_z$ is also estimator $\hat{\mu}_x$ based on μ_x.

The variance of $\hat{\mu}_x$ [62] is

$$\begin{aligned} \operatorname{Var}(\hat{\mu}_x) =& \frac{1}{n}\operatorname{Var}(\bar{Z}) = \frac{1}{n}\Big[E(Z_i)^2 - E(Z_i^2)\Big] \\ =& \frac{1}{n}\Big[E[S_i^2]E[X_i^2] + E[X_i^2] - E(Z_i)^2\Big] \\ =& \frac{1}{n}\Big[\psi(1+\gamma^2)(\sigma_x^2+\mu_x^2) + (1-\psi)(\sigma_x^2+\mu_x^2) - \mu_x^2\Big] \\ =& \frac{1}{n}[\sigma_x^2 + \psi\gamma^2(\sigma_x^2+\mu_x^2)] \end{aligned} \tag{1.22}$$

The additive method (AM) is an extension of the multiplicative technique. Gupta et al. [67] assume that a sample of size n is divided into two sub-samples of sizes n_1 and n_2, where $n_1 + n_2 = n$. In this method, the response to the sensitive question X is a random quantity with unknown mean μ_x (which must be estimated) and unknown variance σ_x^2 where $i = 1, ..., n$.

Let assume that S_j have random quantities of scramble the responses in the sub-sample j for $j = 1, 2$, where θ_j and $\sigma_{s_j}^2$ are the known mean and variance of S_j and $\theta_1 \neq \theta_2$. Suppose that X and S_j have independent random quantities. If a respondent considers the question sensitive, the respondent uses an additive scrambled response; another respondent does not consider the question sensitive and gives their true responses. We assume that the sensitivity level ψ is known. The reported response Z_j in the sub-sample j is:

$$Z_j = \begin{cases} X & \text{with probability} \quad (1-\psi) \\ X + S_j & \text{with probability} \quad \psi \end{cases} \tag{1.23}$$

where $j = 1, 2$. The expected value of Z_j is

$$\begin{aligned} E(Z_j) &= (1-\psi)E(X) + \psi E(S_j + X) \\ &= (1-\psi)\mu_x + \psi(\theta_j + \mu_x) \\ &= \mu_x + \psi\theta_j \end{aligned} \tag{1.24}$$

where $E(S_j) = \theta_j$ for $j = 1, 2$. We estimate $E(Z_j)$ by Z_j that leads to:

$$\hat{\mu}_x = \frac{\theta_2 \bar{Z}_1 - \theta_1 \bar{Z}_2}{\theta_2 - \theta_1} \tag{1.25}$$

where $\bar{Z}_j$ is the sample mean of the responses in the sub-sample j. The values of θ_1 and θ_2 should not be set too near to each other because tiny differences in their values will result in inaccurate variance measurement.

The variance of the estimator $\hat{\mu}_x$ is

$$\begin{aligned} \text{Var}(\hat{\mu}_x) &= \text{Var}\left(\frac{\theta_2 \bar{Z}_1 - \theta_1 \bar{Z}_2}{\theta_2 - \theta_1}\right) \\ &= \frac{1}{(\theta_2 - \theta_1)^2}\left[\frac{\theta_2^2\gamma_1^2}{n_1} + \frac{\theta_1^2\gamma_2^2}{n_2}\right] \end{aligned} \tag{1.26}$$

where the variance of each Z_j is denoted by γ_j^2 and is derived as follows.

$$\begin{aligned} \gamma_j^2 =& (1-\psi)E(X^2) + \psi E(X + S_j)^2 - (E(Z_j))^2 \\ =& (1-\psi)(\sigma_x^2 + \mu_x^2) + \psi E(X^2 + XS_j + S_j^2)^2 - (\mu_x + \psi\theta_j)^2 \\ =& (1-\psi)(\sigma_x^2 + \mu_x^2) + \psi(\sigma_x^2 + \mu_x^2) + 2\psi\mu_x\theta_j + \psi(\theta_j^2 + \sigma_{s_j}^2) - (\mu_x^2 + 2\mu_x\psi\theta_j + \psi^2\theta_j^2) \\ =& (\sigma_x^2 + \psi\sigma_{s_j}^2) + \theta_j^2\psi(1-\psi) \end{aligned} \tag{1.27}$$

and the optimal values of n_1 and n_2 to reduce variance of the estimator $\hat{\mu}_x$ are

$$n_1 = \frac{n\gamma_1\theta_2}{\gamma_1\theta_2 + \gamma_2\theta_1} \tag{1.28}$$

$$n_2 = \frac{n\gamma_2\theta_1}{\gamma_1\theta_2 + \gamma_2\theta_1} \tag{1.29}$$

1.2.3 RRT efficiency comparison

The importance of randomised response designs as a tool for studying sensitive issues increases when they become more effective. We can slightly higher compared the efficiency of randomised response methods by using the optimal design parameters and appropriate sample size. In practice, such measures can help researchers to select good RRT approach in terms of efficiency.

Young et al. [99] use the percent relative efficiency (PRE) to compare the RRT method's efficiency of binary RRT methods. It is defined as the ratio of the theoretical mean square error (MSE) of the estimator of the first RRT method to the mean square error (MSE) of the estimator of the second RRT method. Assume that we have any two RRT methods which have the two estimators of the proportion $\hat{\pi}_{A_1}$ and $\hat{\pi}_{A_2}$ respectively. Then the first RRT method is more efficient than the second RRT method if $\text{MSE}(\hat{\pi}_{A_1}) < \text{MSE}(\hat{\pi}_{A_2})$. Young et al. [99] use the percent relative efficiency (PRE), where

$$\text{PRE} = \frac{\text{MSE}(\hat{\pi}_{A_1})}{\text{MSE}(\hat{\pi}_{A_2})} \tag{1.30}$$

If PRE is greater than 1, we prefer the second method over the first method.

Greenberg [60] defined the efficiency of RRT methods as the ratio of the variance of the estimator $\hat{\pi}_{A_1}$ of the first RRT method and $\hat{\pi}_{A_2}$ of the second method as follows.

$$\text{Efficiency} = \frac{\text{Var}(\hat{\pi}_{A_1})}{\text{Var}(\hat{\pi}_{A_2})} \tag{1.31}$$

if this value is less than 1 that means that the first RRT method is preferred to use instead of the second RRT method. Therefore, lower variance leads to higher efficiency [100].

Similarly, we can compare between the quantitative RRT methods using the variance of the estimators of the two RRT methods. So, the lower variance of the estimator leads to better RRT methods e.g. $\text{Var}(\hat{\mu}_{x_1}) < \text{Var}(\hat{\mu}_{x_2})$ that means the first RRT method is better than the second method [100].

1.2.4 RRT privacy

A fundamental challenge in RRT is how to provide accurate estimates of the population proportion of people with sensitive characteristics while maintaining respondents' privacy. Therefore, several privacy measures Δ have been proposed for qualitative and quantitative RRT, with different implications for optimal method design. These measures typically involve conditional probabilities of the event that the respondents have the sensitive characteristic A (or do not have the sensitive characteristic $\bar{A}$) given the response 'Yes' or 'No' respectively [9, 76, 79].

Zhimin and Zaizai [100] presented the qualitative RRT method for determining the measurement of privacy using the 'Yes' (say, $\dot{Y}$) or 'N' (say, $\dot{N}$) dichotomous response method. To derive the privacy measurement, assume that the conditional probabilities are determined by randomisation device as follows:

$$P[\dot{Y}|\ A] = 1 - P[\dot{N}|\ A] \tag{1.32}$$

$$P[\dot{Y}|\ \bar{A}] = 1 - P[\dot{N}|\ \bar{A}] \tag{1.33}$$

Therefore, the privacy measurement Δ is:

$$\Delta = \left| 1 - \frac{1}{2}\left(\frac{P[\dot{Y}|\ A]}{P[\dot{Y}|\ \bar{A}]} + \frac{P[\dot{N}|\ A]}{P[\dot{N}|\ \bar{A}]} \right) \right| \tag{1.34}$$

where, small values of Δ indicate a high privacy level because the conditional probabilities of the event that the respondents have the sensitive characteristic A (or do not have the sensitive characteristic $\bar{A}$) given the response 'Yes' or 'No' closes to the proportion π_A, and then $\frac{P[\dot{Y}|\ A]}{P[\dot{Y}|\ \bar{A}]}$ and $\frac{P[\dot{N}|\ A]}{P[\dot{N}|\ \bar{A}]}$ closes from 1, then the privacy measurement closes from 0.

To derive the privacy measurement of GB as explained in Section 1.2.1 using Equation (1.34), assume that the conditional probabilities are determined as follows:

$$P[\dot{Y}|\ A] = \pi_B + (1 - \pi_B)\gamma \tag{1.35}$$

$$P[\dot{Y}|\ \bar{A}] = \pi_B(1 - \gamma) \tag{1.36}$$

Then, the privacy measurement of GB is

$$\begin{aligned}\Delta_{GB} &= \left|1-\frac{1}{2}\left(\frac{(\pi_B+(1-\pi_B)\gamma)}{\pi_B(1-\gamma)}+\frac{1-(\pi_B+(1-\pi_B)\gamma)}{1-(\pi_B(1-\gamma))}\right)\right| \\ &= \left|\frac{\gamma(1-2\pi_B(1-\gamma))}{(2\pi_B(1-\gamma)(1-\pi_B(1-\gamma))}\right| \qquad (1.37)\end{aligned}$$

Similarly, the privacy degree of the forced method as explained in Section 1.2.1 is derived using the conditional probabilities of the event that the respondents have the sensitive characteristic A (or do not have the sensitive characteristic $\bar{A}$) given the response 'Yes' or 'No' as follows

$$P[\dot{Y}|\ A] = 1-\gamma_2 \qquad (1.38)$$

$$P[\dot{Y}|\ \bar{A}] = \gamma_1 \qquad (1.39)$$

Then, the corresponding privacy measurement of FM is

$$\begin{aligned}\Delta_{FM} &= \left|1-\frac{1}{2}\left(\frac{1-\gamma_2}{\gamma_1}+\frac{\gamma_2}{1-\gamma_1}\right)\right| \\ &= \left|\frac{\gamma_1(3-2\gamma_1)+\gamma_2(1-2\gamma_1)-1}{2\gamma_1(1-\gamma_1)}\right| \qquad (1.40)\end{aligned}$$

As privacy measure quantitative RRT expectation of the squared of the difference between the reported response Z_i and true response X_i of the sensitive question has been used [100]. The privacy of the quantitative RRT method is

$$\Delta = E(Z_i - X_i)^2 \qquad (1.41)$$

If Δ is larger, the RRT method has greater privacy protection. If a method does not provide any privacy, then $\Delta = 0$. A larger value of Δ of the quantitative RRT method leads to a lower variance of the reported responses Z_i that leads to a higher level of efficiency.

The privacy measure of the Greenberg method Δ_{GM} [100] is

$$\begin{aligned}\Delta_{GM} &= (1-\gamma)E(Y_i-X_i)^2 = (1-\gamma)E(Y_i^2-2X_iY_i+X_i^2) \\ &= (1-\gamma)\left[E(Y_i^2)-2E(X_i)E(Y_i)+E(X_i)^2\right] \\ &= (1-\gamma)\left[(\sigma_y^2+\mu_y^2)-2\mu_y\mu_x+(\sigma_x^2+\mu_x^2)\right] \\ &= (1-\gamma)[\sigma_y^2+\sigma_x^2+(\mu_x-\mu_y)^2] \qquad (1.42)\end{aligned}$$

The privacy measure of EH [51] and MM [62] method are

$$\begin{aligned}\Delta_{EH} &= E(Z_i - X_i)^2 = E(\{S_iX_i - X_i\}^2)\\ &= \frac{1}{\theta^2}E(\{S_i - \theta^2\}^2X_i^2) = (\frac{r}{\theta})^2(\sigma^2 + \mu^2)\\ \Delta_{MM} &= E(Z_i - X_i)^2 = \psi E(\{S_iX_i - X_i\}^2)\\ &= \psi E(\{S_i - 1\}^2X_i^2) = \psi(\gamma)^2(\sigma_x^2 + \mu_x^2) \qquad (1.43)\end{aligned}$$

where $E(S_i) = 1$.

The privacy measure of the additive method Δ_{AM} [71] is

$$\begin{aligned}\Delta_{AM} &= E(Z_j - X_i)^2\\ &= \psi\frac{1}{n_1}\sum_{i=1}^{n_1}(Z_1 - X_i)^2 = \psi E(S_1^2) = \psi(\theta_1^2 + \gamma_1^2) \quad \text{for the first sample}\\ &= \psi\frac{1}{n_2}\sum_{i=1}^{n_2}(Z_2 - X_i)^2 = \psi E(S_2^2) = \psi(\theta_2^2 + \gamma_2^2) \quad \text{for the second sample} \qquad (1.44)\end{aligned}$$

As explained in Sections 1.2.3 and 1.2.4, two crucial factors to take into account, when contrasting any randomised response techniques, which are efficacy level and respondent privacy degree. A technique that offers less privacy has a higher level of efficiency. Conversely, a technique's efficacy will be lower if the level of privacy is higher.

1.3 Reproducibility

Reproducibility of research results is important in many research areas, including science, society, and others. The probability to reproduce the same results of an original experiment in a future experiment using the same computational process, under the same conditions, and with the same study population is referred to as reproducibility. However, there is a reproducibility crisis with contradictory results between initial experiments and subsequent replications due to the fact that many scientific study findings are difficult to interpret or impossible, which affects the validity of the hypotheses they support.

Goodman [59] discovered the importance of statistical test reproducibility to address some common misunderstandings about the statistical p-value. The reproducibility probability, according to Goodman, can be used to show that the p-value could misrepresent the strength of the evidence supporting the null hypothesis. Reproducibility probability can be used to show that the p-value could overstate the strength of the evidence rejecting the null hypothesis.

Therefore, the reproducibility probability for a test is defined as the probability that, if the test is repeated under the same circumstances as the original experiment, the test result, that is, whether the null hypothesis is rejected or not, will be the same.

Senn [89] emphasised the differences between the nature of reproducibility probability and the p-value in the discussion of Goodman's study. Senn agreed with Goodman [59] on the importance of test outcome reproducibility and the probability of reproducibility (RP). Senn [89], on the other hand, disagreed with Goodman's claim that p-values misrepresent evidence against the null hypothesis, emphasising the natural relationship between the p-value and the reproducibility probability where smaller p-value, which measures the strength of the statistical conclusion, leads to larger RP in which is expected to be in the case of a rejected null hypothesis.

According to Goodman [59] and Senn [89] on calculating reproducibility probability (RP), there is a clear recommendation that if a test statistics distribution shows that it is virtually symmetrical under a null hypothesis, then the reproducibility probability is roughly 0.5. This suggestion happened if the test statistic is close from the threshold value. There is a heuristic argument that if the distribution under the null hypothesis of the test statistic is (about) symmetric, then a worst-case scenario would yield an RP of about 0.5 [59, 89]. This is due to the chance that the test statistic value could be equal to the test value of the threshold. Without any additional information, one might expect a repeat of the experiment to yield a second value of the test statistic that is equally likely to be larger or smaller than the original value, and thus the same conclusion with a probability of 0.5. Goodman [59] supports this with a Bayesian argument with a non-informative prior. Senn [89] has discussed difficulties with test reproducibility in real life, such as when a repeated test is performed under different circumstances and by a different team of analysts.

The idea of reproducibility probability (RP) for a given clinical trial was established by Shao and Chow [90]. In this study, the second clinical trial will be conducted with the same research procedure in order to determine whether the first trial's clinical findings can be repeated in the second trial. A two-sided alternative hypothesis tested for a positive known constant. Then, RP is calculated of a statistically significant result for the t-test. They suggested that if the first clinical trial's result is strongly significant, a single clinical trial is acceptable. Shao and Chow [90] considered three approaches to studying reproducibility probability: an estimating the power of a future test based on available test data, an approach in which RP is related to a lower confidence bound, and in which RP is related to a higher confidence bound for the power estimate of the second test, and a third approach is a Bayesian approach. They studies RP in these cases where clinical trial evidence firmly supported a different treatment.

De Martini [47] assessed the reproducibility probability of statistically significant results and proposed statistical tests based on the estimation of reproducibility probability for one-sided and two-sided alternative hypotheses. De Martini demonstrated how to use RP estimation to test parametric hypotheses. The power of the test and the lower confidence bound of the power were considered by De Martini as two definitions of the reproducibility probability of statistically significant results. De Capitani and De Martini [44] considered various estimators of reproducibility probability for the Wilcoxon rank sum test. De Capitani and De Martini [42, 43] investigated estimators for several nonparametric tests, such as the sign and Wilcoxon signed rank tests. They concluded that statistical tests contain randomness, and the reproducibility probability can be also estimated. The RP confidence intervals can be also used to compute statistical tests, and it is possible to demonstrate that the RP pointwise estimator's threshold for defining statistical tests comes out to be 0.5. The RP estimates provide suitable interpretations of the results.

Several further contributions to the development of reproducibility probability are worthy of attention. Posavac [87] offers to assess reproducibility probability by comparing the value of a test statistic based on the actual test and the corresponding threshold value. As a result, if a two-sample test is used, the standard error of the difference between the means of the two relevant samples must be estimated. This allows for the assessment of the probability of a statistically significant precise replication.

Bayesian approaches enable researchers to include more information in their findings and make more informed judgments. Killeen [73] developed the reproducibility probability as an alternative to null-hypothesis significance testing and established a relationship between the reproducibility probability and the effect size. Both the effect size and standard p-values can be regarded as measures the amount of an experimental effect and the significance threshold respectively. Killeen [73] proposed that the effect size be reduced by averaging it using a Bayesian technique with a flat distribution as the prior. While Killeen's study [73], also emphasises the uncertainty that arises throughout the study of this issue. He emphasised the explicit prediction of reproducibility probability and thus believes that RP represents an accurate power test.

Lecoutre et al. [77] provide a discussion of Killeen's technique [73], referring to it as a "calibrated Bayesian predictive probability". Lecoutre et al. [77] agree with Killeen's technique [73] and stated that, despite the technique's success in producing satisfactory results, there is still confusion, as mentioned in Killeen's paper [73]. According to Lecoutre et al. [77], the predictive probabilities form a vital part of the statistical approach and must be taken into consideration. We agree with this. We investigated the variability in RP for estimates using NPI-B and a simple random sampling method to predict the RP for estimates based on the future samples, which produced a variety of results as explained in Chapters 3 and 4.

Miller [83] distinguished between two types of test-repetition circumstances: those in which repetition is carried out by different analysts so that the test circumstance differs from that of the initial experiment, and those in which repetition is carried out by the same researchers as the original experiment and test, under the same circumstance as the original experiment. Miller [83] has doubts about the probability of extracting a valid conclusion from the original experiment because the effect sizes are unknown then the power of this test cannot be determined.

Gelman [57] connected the reproducibility crisis in social science to the default model of constant effects and null hypothesis significance testing, and he claims that Bayesian modelling can lead to a significant increase in understanding social research data. He investigated reproducibility by concentrating on effect sizes, their variability, and the uncertainty in estimating them. The difficulty in fitting interaction models is that they are difficult to estimate to the level of accuracy usually required in practical research and require extra data or prior knowledge. He also mentions that adding further data can well be to make analyses better. Additionally, hierarchical Bayesian analysis can distinguish between large variations in a posterior distribution due to reliable variability in effects across scenarios and large uncertainty due to the non-informativeness of data.

1.4 Nonparametric Predictive Inference(NPI)

Nonparametric Predictive Inference (NPI) is a statistical method based on Hill's assumption $A_{(n)}$ [69], which provides a direct conditional probability for a future observable random quantity based on observed values of related random quantities [13, 23]. NPI can be used for prediction if there is no knowledge of an underlying distribution or if one does not want to use any such knowledge. This can happen if one wants to investigate the hidden impacts of extra structural assumptions underpinning statistical methods. Inferences based on such limited information are also known as low structure or black-box inferences [22].

NPI has been studied for a variety of data types and applications have been presented in statistics, risk and reliability, and operations research. Many studies have proven that NPI has strong statistical features and produces reliable conclusions from predictive inference. NPI for real-valued random values has thus far primarily been limited to a single future observation, but many future observations have been addressed for NPI approaches for statistical process control [10, 11]. It has numerous successful applications in engineering reliability, such as [3, 24, 26, 28, 30]. In addition, NPI has also been employed in the area of finance such as [15, 20, 68].

To introduce the assumption $A_{(n)}$, we have n observed observations $y_{(1)}, ..., y_{(n)}$ and the future observation $m = 1$. Assume that the ordered observed values of the random quantities $Y_1, ..., Y_n$ are denoted by $y_1 < y_2 < ... < y_n$, with the lower bound denoted by

$y_{(0)}$ and the upper bound by $y_{(n+1)}$. It should be noted that $y_{(n+1)}$ is not an observed value for Y_{n+1}. The n observations split the real-line into $n+1$ intervals $I_i = (y_{(i-1)}, y_{(i)})$, where $i = 1, ..., n+1$. The assumption $A_{(n)}$ [69] for one future observation Y_{n+1} is

$$P(Y_{n+1} \in I_i) = \frac{1}{n+1} \quad \text{for} \quad i = 1, ..., n+1 \tag{1.45}$$

The $A_{(n)}$ is a post-data assumption related to exchangeability that makes no further assumptions.

The lower and upper probabilities for any set $\mathfrak{A} \subset \mathbb{R}$ are [13, 23]:

$$\underline{P}(Y_{n+1} \in \mathfrak{A}) = \sum_{i=1}^{n+1} \mathbf{1}\{I_i \subseteq \mathfrak{A}\} P(Y_{n+1} \in I_i) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{I_i \subseteq \mathfrak{A}\} \tag{1.46}$$

$$\overline{P}(Y_{n+1} \in \mathfrak{A}) = \sum_{i=1}^{n+1} \mathbf{1}\{I_i \cap \mathfrak{A} \neq \emptyset\} P(Y_{n+1} \in I_i) = \frac{1}{n+1} \sum_{i=1}^{n+1} \mathbf{1}\{I_i \cap \mathfrak{A} \neq \emptyset\} \tag{1.47}$$

where $\mathbf{1}\{E\}$ is the indicator function where is equal to 1 if event E happens and 0 otherwise. The NPI lower probability is calculated by counting up the probability masses in which $\mathfrak{A}$ must into consideration. The NPI upper probability is calculated by counting up all of the probability masses that could be in $\mathfrak{A}$.

It is evident from the theory of imprecise probability [13, 94, 96, 97] that bounds provide information about the uncertainty of events caused on by restricted information. For the event A, the precise classical probability is only a specific case of the imprecise probability, when $\underline{P}(A) = \overline{P}(A)$ (i. e. the point probability case). However, the situation where $\underline{P}(A) = 0$ and $\overline{P}(A) = 1$ denotes a complete lack of information regarding the event A. We briefly discuss some of the theories of imprecise probability as relevant to $A_{(n)}$-based inference [13]. In general, in imprecise probability theory, $0 \leq \underline{P}(A) \leq \overline{P}(A) \leq 1$, the lower and upper probabilities are conjugated i. e. $\underline{P}(A) = 1 - \overline{P}(A^c)$, where A^c is the complementary event of A, and $\underline{P}(.)$ is super-additive and $\overline{P}(.)$ is sub-additive.

For many events of interest, $A_{(n)}$ is insufficient to determine a precise probability. But it provides optimal probability bounds for all events of interest Y_{n+1} that are lower and upper probabilities. However, in NPI, we use De Finetti's Fundamental Theorem of Probability [45] to determine optimal bounds for the probability of an event of interest [13].

They have strong consistency properties in the theory of imprecise probability [94] and interval probability [97].

NPI has been introduced for a number of applications involving a wide range of data kinds. NPI has been presented for Bernoulli data [22, 29], real-valued data [34, 35, 82], right-censored observations [36, 37]. Coolen and Yan [36, 37] proposed a generalisation of $A_{(n)}$ called " right-censoring-$A_{(n)}$" for right censoring data , circular data [23], multinomial data [22, 27], and bivariate data [38].

1.4.1 NPI for multiple future observations

NPI has been also developed for multiple future real-valued observations where we are interested in $m > 1$. Assume that the ordered observed values of the random quantities $Y_1, ..., Y_n$ are denoted by $y_{(1)} < y_{(2)} < ... < y_{(n)}$, with the lower bound denoted by $y_{(0)}$ and the upper bound by $y_{(n+1)}$. It should be noted that $y_{(n+m)}$ is not an observed value for Y_{n+m} for $m > 1$. The n observations split the real-line into $n + 1$ intervals $I_i = (y_{(i-1)}, y_{(i)})$, where $i = 1, ..., n + 1$.

We assume that all the orderings O_j of the future observations m among the original observations n are equally likely as explained in Section 1.4. For the future observations Y_{n+i}, each ordering can be derived from $S_i^j = \#\{Y_{n+i}, i = 1, ..., n\}$ where $j = 1, 2, ..., \binom{n+m}{n}$. We link the data and future observations via Hill's assumption $A_{(n)}$ [70], or more precisely, via consecutive application of $A_{(n)}, A_{(n+1)}, ..., A_{((n+m)-1)}$ which can be considered as a post-data version of a finite exchangeability assumption for $n + m$ random quantities that are $Y_{n+1}, ..., Y_{n+m}$. A practical interpretation of the $A_{(n)}$ assumptions implies that all possible orderings of n data observations and n future observations are equally likely, where the n data observations and m future observations cannot be separated from one another.

Based on the $A_{(n)}$ assumptions, Equation (1.48) derive the probability of each ordering [34] as follows.

$$P\left(\bigcap_{i=1}^{n+1}\{S_i^j = s_i^j\}\right) = P(O_j) = \binom{n+m}{n}^{-1} \tag{1.48}$$

where the s_i^j are non-negative integers with $\sum_{i=1}^{n+1} s_i^j = m$.

The $A_{(n)}$ assumptions suggest that one has no knowledge of whether particular values of near revealed observations make it more or less likely that future observation will fall between them. Equation (1.48) implies that for each event involving the m future observations, we can count the number of such orderings for which this event holds. In NPI, generally, as described in Section 1.4.1 states that the upper probability of an event is determined by counting all orderings for which it can hold, whereas the lower probability is determined by counting all orderings for which it must hold [13,23]. Several publications are introduced using NPI of reproducibility for multiple future real-valued observations [8,31,91] as explained in Section 1.5.

1.4.2 NPI for Bernoulli random quantities

This section explains NPI for Bernoulli random quantities [22, 29] is one NPI application that is based on a latent variable representation of Bernoulli data. This presentation assumes underlying real-valued quantities and threshold values, so that values to one side of the threshold are successes and values to the other side of the threshold value are failures. The assumption of $A_{(n)}$ yields lower and upper probabilities for the number of successes $A_{(n)}, .., A_{(n+m-1)}$ for m future trials, depending on the number of successes in n observations.

Assume there is a sequence of $n+m$ exchangeable Bernoulli trials, each having the outcomes 'success' and 'failure', with data consisting of s successes in n trials. If Y denotes the random number of successes in trials ranging from 1 to n, then an adequate representation of the data for NPI is $Y_1^n = s$, because all trials are assumed to be exchangeable. Let Y_{n+1}^{n+m} denote the random number of successes in the future trials $n+1$ to $n+m$. Let $R_t = \{r_1, r_2, ..., r_t\}$ with $1 \leq t \leq m+1$ and $0 \leq r_1 < r_2 < ..., r_t \leq m$. The NPI upper probability [22, 29] for the event $Y_{n+1}^{n+m} \in R_t$ given $Y_{n+1}^{n+m} = s$ for $s \in \{0, 1, ..., n\}$ is

$$\overline{P}(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = s) = \binom{m+n}{n}^{-1} \sum_{j=1}^{t} \left[\binom{s-r_j}{s} - \binom{s-r_{j-1}}{s} \right] \binom{n-s+m-r_j}{n-s} \quad (1.49)$$

The corresponding NPI lower probability can be derived using the conjugate property, that is $\underline{P}(A) = 1 - \overline{P}(A^c)$ for any event A and its complementary event A^c.

$$\underline{P}(Y_{n+1}^{n+m} \in R_t \mid Y_1^n = s) = 1 - \overline{P}(Y_{n+1}^{n+m} \in R_t^c \mid Y_1^n = s) \tag{1.50}$$

where R_t^c is the complement of R_t , $R_t^c = \{0, 1, ..., m\}/R_t$.

1.5 NPI for reproducibility

An important characteristic of the practical application of test results is a test's reproducibility. The reproducibility probability (RP), which its definition and interpretation as well as its estimate are not fully specified in the traditional frequentist statistical framework, has attracted a lot of interest recently. The NPI method of frequentist statistics explicitly focuses on future observations while making few assumptions and using lower and upper probabilities to quantify uncertainty. This makes it possible to reach inferences about RP logically given the explicitly predictive nature of NPI.

NPI for reproducibility is first established by Coolen and Bin Himd [22], denoted by NPI-RP, and defined as the probability that, if a test repeated based on an experiment performed in the same way as the original experiment, the test outcome, that is, whether the null hypothesis is rejected or not, would be the same. This was taking into account a few basic nonparametric tests, including the sign test, Wilcoxon's signed rank test, and the two sample rank sum test [58]. NPI for Bernoulli quantities [18], for real-valued data [5] were used for these inferences. This led to NPI lower and upper reproducibility probabilities, denoted by $\underline{RP}$ and $\overline{RP}$, respectively, rather than precisely determined reproducibility probabilities. The NPI lower and upper probability for test reproducibility were calculated for various tests using statistical methods. The NPI-B method, as developed and demonstrated by Bin Himd [33] for the Kolmogorov-Smirnov test, can be used to provide NPI for more complicated test scenarios.

The NPI-RP method is presented for two basic tests using order statistics: a test for a specific population quantile value and a precedence test for comparing data from two populations. These tests are typically used for lifetime data experiments when one wishes to reach a conclusion before all observations are available. For these inferences, NPI for future order statistics is used to provide the lower and upper reproducibility probability for quantile and basic precedence test [7].

Simkus et al. [91] provide an NPI algorithm to assess the reproducibility of the t-test and then use simulations to investigate the reproducibility both under the null and alternative hypotheses. The procedure is to apply NPI reproducibility to real-life applications of a clinical experiment that involves numerous pairwise comparisons of test groups and varying drug concentrations for each group [91].

The nonparametric predictive inference approach for reproducibility of likelihood ratio tests [81]. The idea of this research is to investigate tests between two simple hypotheses on the mean value. The result reveals an upward trend in both the lower and higher reproducibility as well as a distance between the observed likelihood ratios and quantiles.

Coolen and Marques [31] investigated sampling of future data orderings among observed data to obtain the approximate lower and upper reproducibility probability. A new sampling methodology is proposed to overcome the limitations of the usual sampling of orderings method to address scenarios with larger sample sizes [31].

Further work on NPI for reproducibility of statistical inferences based on randomised response data is developed in this thesis. The nature of this thesis is primarily theoretical, with the implementation of the established methodologies shown by example applications.

It is important to mention that there are different methods which are used to compute NPI for reproducibility which is the NPI-B and sampling of ordering methods as discussed in the following section.

1.6 NPI-bootstrap

The key to statistical inference is quantifying the variability of a sample estimate. Making inferences and assuming a probability model are both possible in simple situations, but they can be tricky in complicated ones and can lead to misleading conclusions if the method assumptions are not correct. Efron [50] created a bootstrap method that makes fewer assumptions but requires more computations in order to see through this issue. Due to its ease of use and ability to provide accurate approximations to the sample estimates, there has been an increase in the use of this method. Nonparametric predictive inference bootstrap (NPI-B) is one of the bootstrap methods which is a computational implementation of NPI and it is used to quantify uncertainty in the statistical inference. The original investigation of the NPI-B method was introduced by BinHimd and Coolen [18, 33].

We discuss the performance of the NPI-B method of m future observations where the NPI-B method sample the observations samples values from the data sets and from the interval among them and add them to the data set.

Assume that the ordered observed values of the random quantities $Y_1, ..., Y_n$ are denoted by $y_{(1)} < y_{(2)} < ... < y_{(n)}$, with the lower bound denoted by $y_{(0)}$ and the upper bound by $y_{(n+1)}$. The n observations split the real-line into $n + 1$ intervals $I_i = (y_{(i-1)}, y_{(i)})$, where $i = 1, ..., n + 1$. The assumption

NPI-bootstrap (NPI-B) is based on constructing $n + 1$ intervals from n observations. As in $A_{(n)}$, we create intervals I_i between the observations n where $i = 1, ..., n + 1$, then draw one value from these intervals and add it to the dataset, and then sampling $m - 1$ more values to produce a new sample called an NPI-B sample [18]. All possible orderings of the new observations among the past observations are equally likely to appear in NPI-B. The NPI-B algorithm [18] for one-dimensional real-valued data on a finite interval is as follows:

- Assume there is a data set of n real-valued, one-dimensional observations on a finite interval.
- The partitions $n + 1$ created by n observations.
- Chooses one of the $n + 1$ intervals at random, with equal probability for each interval. From this chosen interval, choose one future value uniformly.
- Increase n to $n + 1$ and add that future value to the data. Steps 2-4 must be repeated with $n + 1$ data to obtain a further future value.

- Repeat this to produce m NPI bootstrap samples $b_1, b_2, ..., b_m$.
- Repeat all of these steps n_B times, to obtain a total of n_B NPI-B samples of size m.

The NPI-Bootstrap method is used in Chapters 3 and 4.

1.7 Sampling of orderings method

Sampling of orderings method (SOM) is a solution when the number of orderings O_j is large of m future observations among large n data observations. Marques et al. [81] illustrate the original work of the SOM method.

To use SOM method based on NPI, we consider that each order that is chosen to be included in the sample must have the same possibility of being selected, and the ordering selection should be independent of the other selections. It is important to note that if the sample size n or the value of orderings sampled is large, the total number of orderings becomes large enough to ignore any potential differences between sampling with or without replacement of these orderings.

To explain SOM method, we need to choose such vectors at random of the orderings $r_1, ..., r_n$ with $r_1 \geq 1$ and $r_{l-1} < r_l$ where $r_n \leq 2n$ for all $l = 2, ..., n$. Take the rank of the l-th ordered data observation among the $2n$ combined data and future observations to be r_l. Then, the future observation data S_l^j is specified as $S_l^j = (r_l - r_{l-1}) - 1$ where $l = 1, ..., n+1$, such that $r_0 = 0$ and $r_{n+1} = 2n + 1$.

This method is used in this thesis in reproducibility for an estimate using the representative sample to generate unlimited orderings for the future observations among the original data as explained in Sections 3.3 and 4.3.

1.8 Thesis outline

The purpose of this thesis is to investigate reproducibility of statistical inferences based on RRT data. This thesis is structured as follows. Chapter 2 considers one-sided and two-sided hypothesis tests based on RRT data. We present a new measure based on the lower and upper reproducibility probability. This work was presented online at International Conference on Advances in Interdisciplinary Statistics and Combinatorics Conference (AISC) in October 2020. Chapter 3 introduces $\epsilon-$reproducibility of an estimate in the general statistical scenario. This work was presented online at the 6^{th} Canadian Conference on Applied Statistics in 2021. Chapter 4 applies the methodologies presented in Chapter 3 to scenarios with data generating from RRT. In Chapter 5, we draw some conclusions and discuss related research challenges.

Chapter 2

Reproducibility of hypothesis tests based on randomised response data

2.1 Introduction

The reproducibility of statistical tests is one of the most important topics in applied statistics, as it has been observed that the conclusions of statistical test could differ if the test is repeated [12].

Some social studies begin with a sensitive question with the aim of eliciting a truth response for the sensitive question of interest with maintaining the respondents' privacy. This method is called randomised response technique. Use of these techniques can be an effective method to quantify sensitive population characteristics.

We are interested in the question if the test were repeated under the same circumstance and with the same sample size, would the same conclusion be reached which is rejection or non-rejection of the null hypothesis?

In this chapter, the reproducibility of statistical tests based on RRT data is discussed, which uses nonparametric predictive inference to predict the results of future hypothesis tests. It compares two RRTs in terms of the reproducibility probability of statistical test or the estimation (variance), and the degree of privacy. It also includes a measure of lower and upper reproducibility of tests based on RRT data. The probability of sensitive questions and the required sample size play an important role in achieving higher reproducibility.

This chapter is organised as follows. Section 2.2 explains NPI for Bernoulli random quantities. Sections 2.3 and 2.4 study the NPI-RP approach for one-sided and two-sided hypothesis tests respectively. Section 2.5 introduces a measure of reproducibility probability of the area under the curve. Section 2.6 calculates the area under MRP of statistical tests based on RRT data. Section 2.7 presents the lower and upper threshold values. Section 2.8 presents a comparison of the reproducibility of statistical tests based on RRT data. Section 2.9 presents a discussion of related topics for further research.

2.2 NPI reproducibility probability for statistical hypothesis tests using Bernoulli data

This section reviews the NPI reproducibility probability for statistical hypothesis tests (NPI-RP) based on Bernoulli data [18, 33]. We use the NPI method as explained in Section 1.4.2 to derive the lower and upper probabilities for the event of interest $Y_{n+1}^{2n} \in I_i$ where $i = 1, ..., n+1$.

It is important to note that the NPI reproducibility probability seems to be predictive in nature, based on data from the first test, one could be able to predict the results of a future test assuming it would have the same sample size and become performed under similar circumstances. Therefore, we assume that the sample size of the original sample size n is equal to the future sample n.

We suppose that a sequence of $2n$ exchangeable Bernoulli trials, each with ‘Yes’ and ‘No’ values. Let Y_1^n denote the random number of ‘Yes’ answers in trials 1 to n and Y_{n+1}^{2n} denote the random number of ‘Yes’ answers in trials $n+1$ to $2n$. Based on the basic method represented by Coolen [22] and Coolen and Coolen-Schrijner proposal [28], the NPI lower and upper probability for the events $Y_{n+1}^{2n} \geq C$ are derived as follows.

$$\underline{P}(Y^{2n}_{n+1} \geq C \mid Y^n_1 = y) = 1 - \binom{2n}{n}^{-1} \times \left[\sum_{l=1}^{C-1} \binom{y+l-1}{y-1} \binom{2n-y-l}{n-y} \right] \tag{2.1}$$

and

$$\begin{aligned} \overline{P}(Y^{2n}_{n+1} \geq C \mid Y^n_1 = y) = \binom{2n}{n}^{-1} \Bigg[\binom{y-C}{y} \binom{2n-y-C}{n-y} \\ + \sum_{y=C+1}^{n} \binom{y-l-1}{y-1} \binom{2n-y-l}{n-y} \Bigg] \end{aligned} \tag{2.2}$$

where C is the rejection threshold, and $y \in \{1, ..., n-1\}$. If the observed data are all 'Yes' answers (so $y = n$), or all 'No' answers (so $y = 0$), then the NPI upper probabilities are

$$\overline{P}(Y^{2n}_{n+1} \geq C \quad \mid Y^n_1 = n) = 1 \tag{2.3}$$

$$\overline{P}(Y^{2n}_{n+1} \geq C \quad \mid Y^n_1 = 0) = \binom{2n}{n}^{-1} \binom{2n-C}{n} \tag{2.4}$$

and NPI lower probabilities are:

$$\underline{P}(Y^{2n}_{n+1} \geq C \quad \mid Y^n_1 = n) = 1 - \left[\binom{2n}{n}^{-1} \binom{n+C-1}{n} \right] \tag{2.5}$$

$$\underline{P}(Y^{2n}_{n+1} \geq C \quad \mid Y^n_1 = 0) = 0 \tag{2.6}$$

This method can be applied to reproducibility of one-sided tests based on RRT data, Section 2.3 introduces more details.

For the two thresholds l, r, the n future random quantities given n observations can be represented with Bernoulli quantities represented by observations on the real line, such that the non-rejection region which includes the 'Yes' answers in the range between the endpoints l and r respectively.

Suppose there is a non-rejection area R between the two points l and r. If the event $Y^{2n}_{n+1} \in \{l, ..., r\}$ where $1 \leq l < r \leq n$, the H_0 is rejected if and only if $y < l$ and $y \geq r$. Then, the related NPI lower and upper probabilities, assuming $Y^n_1 = y$, are easily determined from Equations (2.7) and (2.8), using Coolen's paper [22] as follows.

The NPI upper probability is

$$\overline{P}(Y_{n+1}^{2n} \in \{l, ..., r\} \mid Y_1^n = y) = \binom{2n}{n}^{-1} \times$$
$$\left[\binom{y+r+1}{y}\binom{2n-y-r-1}{n-y} + \sum_{i=l+2}^{r-1}\left\{\binom{y+i-1}{y-1}\binom{2n-y-i}{n-y}\right\}\right] \tag{2.7}$$

The NPI lower probability is

$$\underline{P}(Y_{n+1}^{2n} \in \{l, ..., r\} \mid Y_1^n = y) =$$
$$1 - \binom{2n}{n}^{-1} \times \left[\binom{2n-y}{n-y} + \sum_{i=1}^{r-1}\left\{\binom{y+i-1}{y-1}\binom{2n-y-i}{n-y}\right\} + \left\{\binom{y+r+1}{y} - \right.\right.$$
$$\left.\left.\binom{y-l-1}{y}\right\} \times \binom{2n-y-(r+1)}{n-y} + \sum_{i=r+1}^{n}\left\{\binom{y-i-1}{y-1}\binom{2n-y-i}{n-y}\right\}\right] \tag{2.8}$$

This method can be applied to reproducibility of two-sided tests based on RRT data, Section 2.4 introduces more details.

2.3 Reproducibility of one-sided hypothesis tests based on RRT data

Reproducibility of one-sided hypothesis tests based on randomised response data (NPI-RP-RRT) shows how probably it is that a future test of qualitative RRT data will lead to the same conclusion as the original test.

Based on RRT methods, we consider the hypothesis test for the proportion of people with a sensitive characteristic A, where $H_0^{'}$ is the null hypothesis of the proportion $\pi_A = \pi_{A_0}$, and $H_1^{'}$ is the alternative hypothesis of the proportion $\pi_A > \pi_{A_0}$, as follows:

$$H_0' : \pi_A = \pi_{A_0} \quad \text{and} \quad H_1' : \pi_A > \pi_{A_0} \tag{2.9}$$

where $\pi_{A_0} \in [0, 1]$.

Give $P^* = P_0^*$ is the function of the proportion π_{A_0} of people who answer ‘Yes’ to the question of RRT method.

We assume the proportion of people with a sensitive characteristic π_{A_0} and calculate the proportion of people who will say 'Yes' using P^* as explained in Section 1.2.1, to write the corresponding hypothesis test as follows:

$$H_0 : P^* = P_0^* \text{ and } H_1 : P^* > P_0^* \tag{2.10}$$

A logical test rule is rejecting the null hypothesis if $Y_1^n \geq C$ for chosen significance level α is:

$$P(Y_1^n \geq C \mid H_0) \leq \alpha \tag{2.11}$$

The NPI upper and lower reproducibility probabilities for the event $Y_{n+1}^{2n} \geq C$ are expressed as function of y, with respect to C as rejection threshold and α as level of significance using the equations in Section 2.2, as follows

$$\overline{RP}(y) = \overline{P}(Y_{n+1}^{2n} \geq C \mid Y_1^n = y), \qquad \underline{RP}(y) = \underline{P}(Y_{n+1}^{2n} \geq C \mid Y_1^n = y) \tag{2.12}$$

If we observe $Y_1^n < C$, the upper and lower reproducibility probabilities of this event $Y_1^n < C$ using the conjugacy property are as follows:

$$\overline{RP}(y) = 1 - \underline{P}(Y_{n+1}^{2n} \geq C \mid Y_1^n = y), \qquad \underline{RP}(y) = 1 - \overline{P}(Y_{n+1}^{2n} \geq C \mid Y_1^n = y) \tag{2.13}$$

Examples 2.3.1 and 2.3.2 illustrate this method using the GB and the FM methods explained in Section 1.2.1.

Example 2.3.1 This example explains NPI reproducibility for one-sided tests based on data collected using GB method (NPI-RP-GB). Suppose that we have a sample with size $n = 30$ and are interested in a sensitive characteristic A. The unknown proportion of people with the sensitive characteristic is π_{A_0}, and π_B is the proportion of people who would respond 'Yes' to the unrelated question where π_B is known and equal to 0.3. In this example, we assume that a randomisation device is used with a probability that the sensitive question is asked equal to $\gamma = 0.7$.

To start with, we assume that we need to test the null hypothesis that the proportion of people who have the sensitive characteristic π_{A_0} is equal to 0.70, against the alternative hypothesis that is greater than 0.70.

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9956	0.9980	22	0.5	0.6145
1	1.0000	1.0000	12	0.9909	0.9956	23	0.6145	0.7240
2	1.0000	1.0000	13	0.9824	0.9909	24	0.7240	0.8198
3	1.0000	1.0000	14	0.9680	0.9824	25	0.8198	0.8954
4	1.0000	1.0000	15	0.9449	0.9680	26	0.8954	0.9479
5	1.0000	1.0000	16	0.9101	0.9449	27	0.9479	0.9790
6	1.0000	1.0000	17	0.8605	0.9101	28	0.9790	0.9939
7	0.9999	1.0000	18	0.7941	0.8605	29	0.9939	0.9990
8	0.9997	0.9999	19	0.7102	0.7941	30	0.9990	1
9	0.9992	0.9997	20	0.6106	0.7102			
10	0.9980	0.9992	21	0.5	0.6106			

Table 2.1: NPI-RP-GB at $\alpha = 0.05, C = 22$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9981	0.9992	22	0.5	0.6145
1	1.0000	1.0000	12	0.9959	0.9981	23	0.5	0.6195
2	1.0000	1.0000	13	0.9916	0.9959	24	0.6195	0.7340
3	1.0000	1.0000	14	0.9837	0.9916	25	0.7340	0.8333
4	1.0000	1.0000	15	0.9702	0.9837	26	0.8333	0.9097
5	1.0000	1.0000	16	0.9483	0.9702	27	0.9097	0.9601
6	1.0000	1.0000	17	0.9149	0.9483	28	0.9601	0.9872
7	1.0000	1.0000	18	0.8666	0.9149	29	0.9872	0.9977
8	0.9999	1.0000	19	0.8007	0.8666	30	0.9977	1
9	0.9997	0.9999	20	0.7163	0.8007			
10	0.9992	0.9997	21	0.6145	0.7163			

Table 2.2: NPI-RP-GB at $\alpha = 0.01, C = 23$.

So the hypotheses are

$$H_0' : \pi_A = 0.7 \text{ and } H_1' : \pi_A > 0.7 \tag{2.14}$$

and we test with a level of significance $\alpha = 0.05$.

These hypotheses lead to the null and alternative hypotheses for P^* for which the observed number of 'Yes' answers is derived using Equation (1.5) in Section 1.2.1.

$$H_0 : P^* = 0.58 \text{ and } H_1 : P^* > 0.58 \tag{2.15}$$

The corresponding threshold value C for the one-sided test is 22. Therefore, H_0 is rejected if $Y_1^n \geq C$; otherwise, it is not rejected. In this example, the null hypothesis $\pi_A = 0.7$ is not rejected. Then, the p-value can be computed as follows: $P(Y_1^n \geq 23|n =$

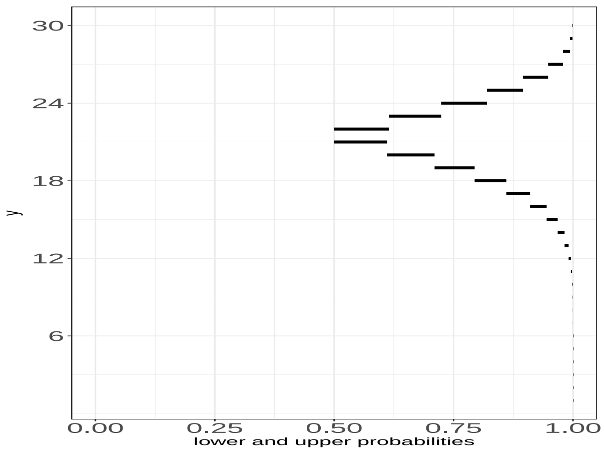


Figure 2.1: NPI-RP of one-sided test based on GB data of $n = 30,\ \pi_{A_0} = 0.7,\ \pi_B = 0.3,$ $\gamma = 0.7,\ \alpha = 0.05,\ P_0^* = 0.58$

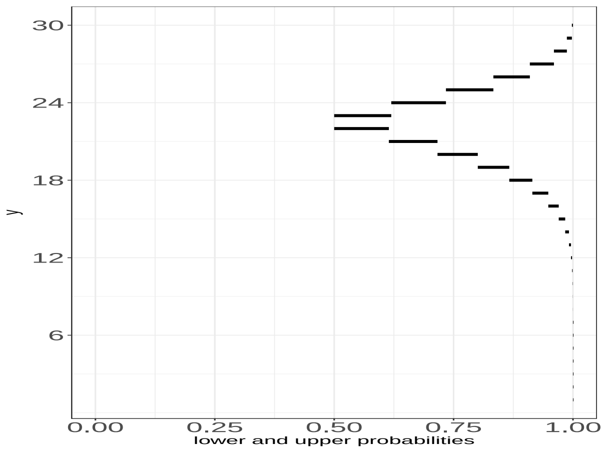


Figure 2.2: NPI-RP of one-sided test based on GB data of $n = 30,\ \pi_{A_0} = 0.7,\ \pi_B = 0.3,$ $\gamma = 0.7,\ \alpha = 0.01,\ P_0^* = 0.58$

$30, P_0^* = 0.58) = 0.0296$ which is less than 0.05. Therefore, we conclude that H_0 can be rejected if $Y_1^n \leq C$. Then, the claim that the proportion of people who answer 'Yes' is greater than 0.7 would be not rejected, at the 0.05 significance level.

The NPI lower and upper reproducibilities probability for the event $Y_{n+1}^{2n} \geq C$ is derived from Equation (2.12) whereas the NPI lower and upper reproducibility probabilities for the event $Y_{n+1}^{2n} < C$ can be derived from Equation2.13.

The NPI lower and upper reproducibility probabilities of the event $Y_{n+1}^{2n} \geq C = 22$ given H_0 are presented in Tables 2.1 and 2.2. For all responses which are 'Yes' or 'No' for the value $y = 0$ or $y = n$ which occurs if all observations in the original test are 'Yes' answers, or if all are 'No' answers respectively. If all of the responses are 'Yes' ('No'), the data has no effect on the probability that 'No' ('Yes') responses will never happen.

The minimum value of the lower reproducibility probability is 0.5 which means the test statistic is symmetric, then a worst-case scenario would result in RP of 0.5 [59, 89].

The maximum value of the upper reproducibility probability is 1. This occurs when all observations in the original test are greater than C, so $y = n$, in which case H_0 is rejected, at any level of significance α; this shows that the possibility that no future observations will exceed the value cannot be excluded with no evidence in the original data to indicate that the data values can exceed the rejection threshold. Note that the corresponding NPI lower reproducibility probability is less than 1 for $y = 0$, reflecting that the original data set provides only limited information which leads to an increase in the NPI lower reproducibility probability towards 1.

If the original test does not lead to the rejection of $H'_0 : \pi_A = 0.7$, such that $Y_1^n \leq C = 22$ at $\alpha = 0.05$, then the reproducibility of statistical tests is the probability that the null hypothesis will be not rejected in the future test. The values of $Y_1^n = y$ above the rejection threshold, $C = 22$ leads to non-rejection of H'_0. Then the NPI lower reproducibility of y is equal to the reproducibility probability of $y + 1$ such that $\underline{RP}(y) = \overline{RP}(y + 1)$.

Conversely, if the original test does lead to the rejection of $H'_0 : \pi_A = 0.7$, the reproducibility probability of the y which is greater than the rejection threshold $C = 22$, the NPI-RP of the event $Y_1^n \geq 22$ given $H_0 : P_0^* = 0.58$ produces a different relationship between the lower and upper probabilities of the events: $\underline{RP}(y) = \overline{RP}(y - 1)$.

Figures 2.1 and 2.2 show NPI-RP-GB at significance level $\alpha = 0.05$ and $\alpha = 0.01$, and their rejection threshold values of 23 and 24 respectively. As already mentioned in Section 1.5, the reproducibility probability of statistical tests measures the probability that the same decision would be made if a test were repeated under the same circumstances. The larger value of the lower RP suggests that a test never would be repeated with high probability, and the same decision regarding rejection of the null hypothesis would be reached.

As shown in Figure 2.3 and Table 2.3, there is a special case for NPI-RP-GB when

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.8472	0.9011	22	0.9402	0.9665
1	1.0000	1.0000	12	0.7779	0.8472	23	0.9665	0.9828
2	0.9999	1.0000	13	0.6944	0.7779	24	0.9828	0.9920
3	0.9997	0.9999	14	0.6002	0.6944	25	0.9920	0.9967
4	0.9988	0.9997	15	0.5	0.6002	26	0.9967	0.9988
5	0.9967	0.9988	16	0.5	0.6002	27	0.9988	0.9997
6	0.9920	0.9967	17	0.6002	0.6944	28	0.9997	0.9999
7	0.9828	0.9920	18	0.6944	0.7779	29	0.9999	1.0000
8	0.9665	0.9828	19	0.7779	0.8472	30	1.0000	1
9	0.9402	0.9665	20	0.8472	0.9011			
10	0.9011	0.9402	21	0.9011	0.9402			

Table 2.3: NNPI-RP-GB of $n = 30,\ \pi_{A_0} = 0.6,\ \pi_B = 0.1,\ \gamma = 0.55,\ P_0^* = 0.3750,\ \alpha = 0.05$

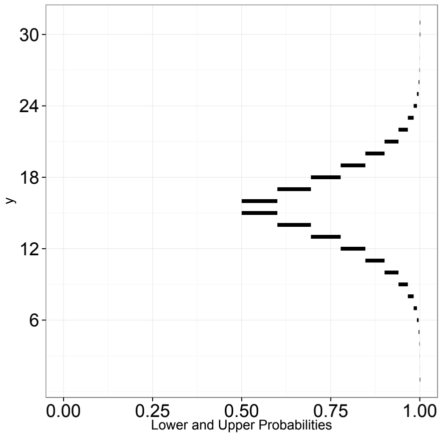


Figure 2.3: NPI-RP-GB of $n = 30,\ \pi_{A_0} = 0.6,\ \pi_B = 0.1,\ \gamma = 0.55,\ P_0^* = 0.3750,\ \alpha = 0.05$

$\pi_{A_0} = 0.6$, $\pi_B = 0$, $n = 30$, $\gamma = 0.55$, $\alpha = 0.05$ and then $P^* = (0.6)(0.55) + (0.45)(0.1) = 0.3750$. The reproducibility of statistical tests based on GB data is symmetric around the rejection threshold $C = \frac{n}{2} = 30$. This means that $\underline{RP}(y = L) = \underline{RP}(y = 30 - L + 1)$ and $\overline{RP}(y = L) = \overline{RP}(y = 30 - L + 1)$, $L = 0, 1, ..., \frac{n}{2}$.

Example 2.3.2 This example introduces the reproducibility probability for one-sided hypothesis tests with data collected using the forced method. Assume that a sample of size n is taken from a population with a possible sensitive characteristic A. Suppose that the proportion of the sensitive characteristic is π_{A_0}. The randomisation device leads to the sensitive question being asked with probability $\gamma = 0.75$, or the answer is forced to 'Yes' with probability $\gamma_1 = 0.10$ or forced to 'No' with probability $\gamma_2 = 0.15$. The significance

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9993	0.9997	22	0.6195	0.7240
1	1.0000	1.0000	12	0.9983	0.9993	23	0.5	0.6195
2	1.0000	1.0000	13	0.9964	0.9983	24	0.5	0.6260
3	1.0000	1.0000	14	0.9925	0.9964	25	0.6260	0.7469
4	1.0000	1.0000	15	0.9854	0.9925	26	0.7469	0.8505
5	1.0000	1.0000	16	0.9731	0.9854	27	0.8505	0.9273
6	1.0000	1.0000	17	0.9527	0.9731	28	0.9273	0.9738
7	1.0000	1.0000	18	0.9210	0.9527	29	0.9738	0.9947
8	1.0000	1.0000	19	0.8742	0.9210	30	0.9947	1
9	0.9999	1.0000	20	0.8092	0.8742			
10	0.9997	0.9999	21	0.7240	0.8092			

Table 2.4: NPI-RP-FM of $\alpha = 0.05,\ \pi_{A_0} = 0.7,\ \gamma_1 = 0.15,\ \gamma_2 = 0.10,\ C = 24.$

level for the hypothesis test is $\alpha = 0.05$.

To start with, assume a sample with size $n = 30$, the null hypothesis that the proportion of people who have characteristic is $\pi_{A_0} = 0.70$, which is tested against $\pi_{A_0} > 0.70$. So the hypothesis test is

$$H_0' : \pi_A = 0.7 \ \text{ vs } \ H_1' : \pi_A > 0.7 \tag{2.16}$$

which is corresponding to the test:

$$H_0 : P^* = 0.625 \ \text{ vs } \ H_1 : P^* > 0.625 \tag{2.17}$$

Using the probability P^* of respondents who say 'Yes' in Equation (1.8), then

$$P_0^* = \gamma_1 + \pi_{A_0}(1 - \gamma_1 - \gamma_2) = 0.625 \tag{2.18}$$

The threshold value C for the one-sided test is $C = 24$. Therefore, H_0 is rejected if $Y_1^n \geq C$; otherwise, it is not rejected. In this example, the null hypothesis $\pi_A = 0.7$ is rejected. Then, p-value can be computed as follows: $P(Y_1^n \geq 25|n = 30, P_0^* = 0.625) = 0.0326 < 0.05$. It is concluded that H_0 can be rejected if $Y_1^n \leq C$. Therefore, the claim that the true proportion of people who answer 'Yes' is 0.7 would be rejected, at the 0.05 significance level. The NPI lower and upper probabilities for the event $Y_{n+1}^{2n} \geq C$ are shown in Tables 2.4 and 2.5.

The NPI lower and upper probabilities for the event $Y_{n+1}^{2n} \geq C$ are shown in Tables 2.4 and 2.5. At $\alpha = 0.05$, the rejection threshold is $C = 24$, whereas at $\alpha = 0.01$, the rejection threshold is $C = 25$.

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.9998	0.9999	22	0.7340	0.8198
1	1.0000	1.0000	12	0.9994	0.9998	23	0.6260	0.7340
2	1.0000	1.0000	13	0.9986	0.9994	24	0.5	0.6260
3	1.0000	1.0000	14	0.9969	0.9986	25	0.5	0.6347
4	1.0000	1.0000	15	0.9937	0.9969	26	0.6347	0.7642
5	1.0000	1.0000	16	0.9875	0.9937	27	0.7642	0.8729
6	1.0000	1.0000	17	0.9765	0.9875	28	0.8729	0.9486
7	1.0000	1.0000	18	0.9580	0.9765	29	0.9486	0.9881
8	1.0000	1.0000	19	0.9284	0.9580	30	0.9881	1
9	1.0000	1.0000	20	0.8837	0.9284			
10	0.9999	1.0000	21	0.8198	0.8837			

Table 2.5: NPI-RP-FM of $\alpha = 0.01,\ \pi_{A_0} = 0.7,\ \gamma_1 = 0.15,\ \gamma_2 = 0.10, C = 25$.

For all responses which are ‘Yes’ or ‘No’ for the value $y = 0$ or $y = n$ which occurs if all observations in the original test are ‘Yes’ answers, or if all are ‘No’ answers respectively. If all of the responses are ‘Yes’ (‘No’), the data has no effect on the probability that ‘No’ (‘Yes’) responses will never happen.

The minimum value of the lower reproducibility probability is 0.5 which means the test statistic has worst-case scenario would result in RP of 0.5 [59,89] and according to Sulafah’s proof [18], the lower reproducibility obtains a minimum value of 0.5, and the greater reproducibility probabilities obtain a higher value as the test approaches the threshold C.The maximum value of the upper reproducibility probability is 1. This occurs when all observations in the original test are greater than C, so $y = n$, in which case H_0 is rejected, at any level of significance α.

If the original test does not lead to the rejection of $H_0' : \pi_A = 0.7$, such that $Y_1^n \leq C = 24$ at $\alpha = 0.05$, then the reproducibility of statistical tests is the probability that the null hypothesis will be not rejected in the future test. The value of $Y_1^n = y$ above the rejection threshold, $C = 24$ leads to non-rejection of $H_0' : P^* = 0.625$. Then the NPI lower reproducibility of y is equal to the reproducibility probability of $y + 1$ such that $\underline{RP}(y) = \overline{RP}(y + 1)$.

Conversely, if the original test does lead to the rejection of $H_0' : \pi_A > 0.7$, the reproducibility probability of the y which is greater than the rejection threshold $C = 24$, the NPI-RP of the event $Y_1^n \geq 24$ given H_0 produces a different relationship between the

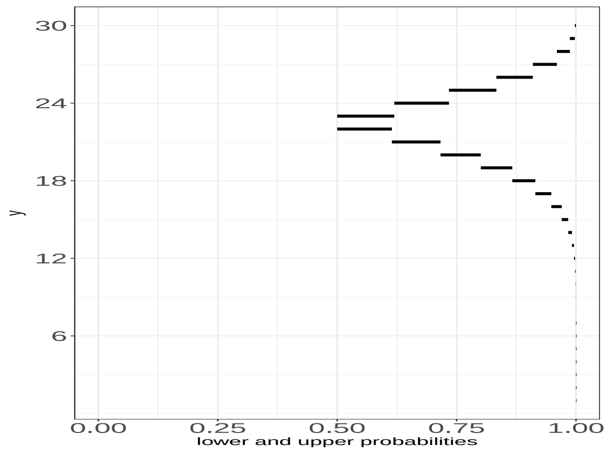


Figure 2.4: NPI-RP of one-sided test based on the forced data of $n = 30$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.10$, $\gamma_2 = 0.15$, $\alpha = 0.05$, $P_0^* = 0.625$

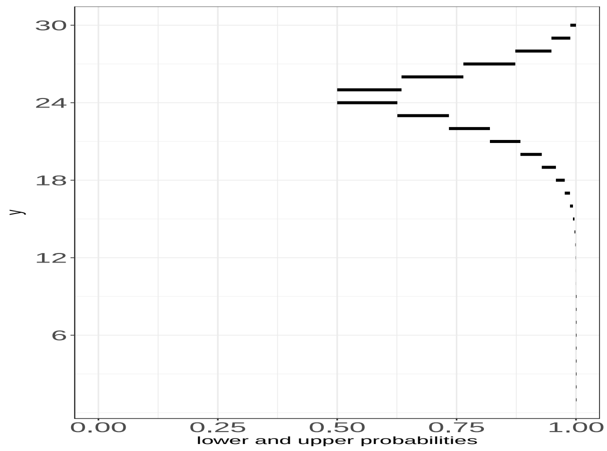


Figure 2.5: NPI-RP of one-sided test based on forced data of of $n = 30$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.10$, $\gamma_2 = 0.15$, $\alpha = 0.01$, $P_0^* = 0.625$

lower and upper probabilities of the events: $\underline{RP}(y) = \overline{RP}(y-1)$.

Figures 2.5 and 2.4 show NPI-RP-FM at significance level $\alpha = 0.05$ and $\alpha = 0.01$, and their rejection threshold values of 24 and 25 respectively. Obviously, for $Y_1^n = y$ such that the null-hypothesis is not rejected, the NPI lower and upper reproducibility probabilities are higher for $\alpha = 0.01$ than for $\alpha = 0.05$, while the opposite is true for $Y_1^n = y$ such that the null-hypothesis is rejected. These properties directly follow from the fact that the null-hypothesis is rejected for fewer values for $Y_1^n = y$ if the significance level α is smaller.

2.4 Reproducibility of two-sided hypothesis tests based on RRT data

In this section, the reproducibility of two-sided hypothesis tests based on randomised response data is studied by deriving the NPI lower and upper reproducibility probabilities. This follows the main steps of the one-sided test represented in Section 2.3, with two rejection regions. The threshold values of a two-sided test splits the rejection region into two regions.

Suppose that there is a sequence of $2n$ exchangeable Bernoulli trials, each with 'Yes' and 'No' as possible responses. We apply NPI as explained in Section 1.5 for random quantity Y_1^n of 'Yes' answer in trials 1 to n of the data $Y_1^n = y$, and random quantity Y_{n+1}^{2n} of 'Yes' answers in trials $n+1$ to $2n$. We assume the future number of observations is equal to the number of data observations n.

To start with, we consider the hypothesis tests for the proportion of people with a sensitive characteristic A, where $H_0^{'}$ is the null hypothesis of the proportion $\pi_A = \pi_{A_0}$ and alternative hypothesis as follows:

$$H_0^{'} : \pi_A = \pi_{A_0} \quad \text{versus} \quad H_1^{'} : \pi_A \neq \pi_{A_0} \tag{2.19}$$

where $\pi_{A_0} \in [0, 1]$ and the level of significance is α.

The null hypothesi $H_0^{'}$ is rejected if $Y_1^n \geq r{+}1$ or $Y_1^n \leq l - 1$, otherwise, H_0' is not rejected, where the events $Y_1^n \leq l-1$ and $Y_1^n \geq r+1$ are derived using Binomial distribution and level of significance α from the following formulas:

$$P(Y_1^n \geq r \ |H_0 \ \text{ is true}) \leq \frac{\alpha}{2} \tag{2.20}$$

$$P(Y_1^n \leq l \ |H_0 \ \text{ is true}) \leq \frac{\alpha}{2} \tag{2.21}$$

We have the alternative hypothesis $H_1' : \pi_A \neq \pi_{A_0}$, such that π_{A_0} is the hypothesised proportion of people who have the sensitive characteristic. This relates to P_0^* as mentioned in Equations (1.5) and (1.8) in Section 1.2.1.

Then, the NPI lower reproducibility probability of the event $Y_{n+1}^{2n} \leq l-1 \wedge Y_{n+1}^{2n} \geq r+1$ if the original test led to rejection of H_0 , given $Y_1^n = y$, is

$$\underline{RP}(y) = \underline{P}(Y_{n+1}^{2n} \leq l-1 \wedge Y_{n+1}^{2n} \geq r+1 \mid Y_1^n = y) = 1 - \overline{P}(Y_{n+1}^{2n} \in \{l-1, ..., r+1\} \mid Y_1^n = y) \tag{2.22}$$

and the corresponding NPI upper reproducibility probability is

$$\overline{RP}(y) = \overline{P}(Y_{n+1}^{2n} \leq l-1 \vee Y_{n+1}^{2n} \geq r+1 \mid Y_1^n = y) = \underline{P}(Y_{n+1}^{2n} \in \{l-1, ..., r+1\} \mid Y_1^n = y) \tag{2.23}$$

If H_0 is not rejected in the original test, then the NPI lower reproducibility probability is

$$\begin{aligned} \underline{RP}(y) &= \underline{P}(Y_{n+1}^{2n} \in \{l-1, ..., r+1\} \mid Y_1^n = y) \\ &= 1 - \overline{P}(Y_{n+1}^{2n} \in \{0, 1, ..., l-1\} \cup \{r+1, ..., n\} \mid Y_1^n = y) \end{aligned} \tag{2.24}$$

and the NPI upper reproducibility probability of the event $Y_{n+1}^{2n} \in \{l-1, .., r+1\}$ is

$$\overline{RP}(y) = \overline{P}(Y_{n+1}^{2n} \in \{l, ..., r\} \mid Y_1^n = y) \tag{2.25}$$

For the case of rejection of the null hypothesis in the original test given either $y = l-1$ or $y = r+1$, the minimum value that can occur for the NPI lower reproducibility probability for this two-sided binomial test is less than 0.5. This is evident from the equations for the NPI lower reproducibility probabilities when $y = l-1$ or $y = r+1$, with $\underline{P}(Y_{n+1}^{2n} \leq l-1 \vee Y_{n+1}^{2n} \geq r+1 \mid Y_1^n = y) < 0.5$ because of the two rejection regions for H_0, the event here differs from the one-sided test, as we now sum up the probability masses for the two events $Y_{n+1}^{2n} \leq l-1$ and $Y_{n+1}^{2n} \geq r+1$ both given $Y_1^n = y$, where the probability for the event $Y_{n+1}^{2n} \leq l-1$ is equal to 0.5 and $\underline{P}(Y_{n+1}^{2n} \geq r+1 \mid Y_1^n = y) > 0$.

If H_0 is not rejected in the original test, and $l-1 \leq y \leq r+1$, then the NPI lower reproducibility probability for the event $Y_{n+1}^{2n} \leq l-1$ given $Y_1^n = y$ would be less than 0.5 and the NPI lower reproducibility probability for the event $Y_{n+1}^{2n} \leq r+1$ given $Y_1^n = y$ is less than 0.5. The NPI reproducibility upper probabilities maximum value that can occur is less than 1 in the scenario where the original test did not reject the null hypothesis. This happens when y closes from $l-1$ or $r+1$.

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	11	0.6058	0.7065	22	0.3764	0.4959
1	0.9998	1.0000	12	0.3872	0.4956	23	0.6162	0.7281
2	0.9989	0.9998	13	0.4824	0.5935	24	0.7246	0.8215
3	0.9963	0.9989	14	0.5706	0.6803	25	0.8200	0.8960
4	0.9898	0.9963	15	0.6428	0.7490	26	0.8954	0.9481
5	0.9765	0.9898	16	0.6912	0.7941	27	0.9479	0.9791
6	0.9527	0.9765	17	0.7099	0.8119	28	0.9791	0.9939
7	0.9149	0.9528	18	0.6960	0.8003	29	0.9939	0.9990
8	0.8607	0.9152	19	0.6501	0.7595	30	0.9990	1
9	0.7893	0.8614	20	0.5761	0.6917			
10	0.7029	0.7910	21	0.4816	0.6015			

Table 2.6: NPI-RP-GB at $\alpha = 0.05, l = 12, r = 22$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.9998	1	11	0.4981	0.6046	22	0.4980	0.6137
1	0.9988	0.9998	12	0.6013	0.7002	23	0.3797	0.4997
2	0.9949	0.9988	13	0.6936	0.7811	24	0.6196	0.7342
3	0.9848	0.9949	14	0.7689	0.8441	25	0.7340	0.8334
4	0.9642	0.9848	15	0.8227	0.8875	26	0.8333	0.9097
5	0.9284	0.9642	16	0.8522	0.9108	27	0.9097	0.9601
6	0.8742	0.9284	17	0.8555	0.9137	28	0.9601	0.9872
7	0.8008	0.8743	18	0.8320	0.8959	29	0.9872	0.9977
8	0.7103	0.8009	19	0.7818	0.8569	30	0.9977	1
9	0.6078	0.7105	20	0.7066	0.7961			
10	0.3915	0.4997	21	0.6099	0.7142			

Table 2.7: NPI-RP-GB at $\alpha = 0.01, l = 10, r = 23$

Section 2.2 explains the derivation of these equations. Furthermore, we illustrate the results in the following examples.

Example 2.4.1 This example illustrate the NPI reproducibility probability of two-sided hypothesis tests from data collected by the Greenberg method. Suppose that we have a sample size of $n = 30$ from a population who have a sensitive characteristic A. Let the proportion of people who have the sensitive characteristic be π_A, and the known proportion of those with the unrelated characteristic is $\pi_B = 0.3$. Assume that we use a randomisation device with probability $\gamma = 0.7$ for the sensitive question being asked.

We want to test the null hypothesis that the probability of the proportion of people who have sensitive characteristics is $\pi_{A_0} = 0.7$ against the alternative hypothesis that $\pi_{A_0} \neq 0.7$. So,

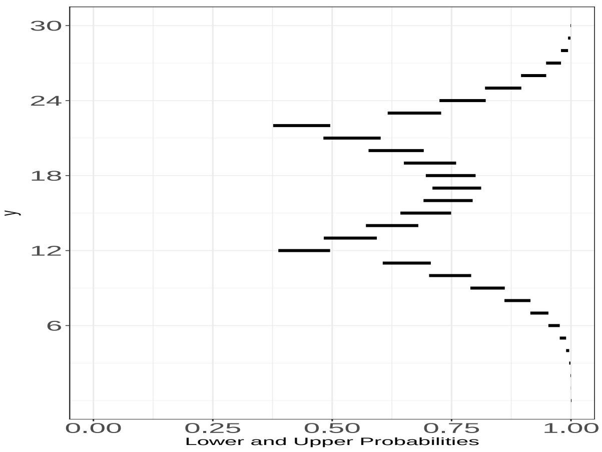


Figure 2.6: NPI-RP of two-sided tests using data collected from the GB method of $n = 30$, $\pi_{A_0} = 0.7, \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P_0^* = 0.58$

$$H_0' : \pi_A = 0.7 \text{ and } H_1' : \pi_A \neq 0.7 \tag{2.26}$$

The corresponding null and alternative hypotheses for P^* using Equation (1.5) for NPI-RP-GB are:

$$H_0 : P^* = 0.58 \text{ and } H_1 : P^* \neq 0.58 \tag{2.27}$$

where the proportion of people saying 'Yes' is $P_0^* = (0.7)(0.7) + (1 - 0.7)(0.3) = 0.58$.

As shown in Figures 2.6 and 2.7 and Tables 2.6 and 2.7, at the significance level $\alpha = 0.05$ and $\alpha = 0.01$ respectively where $H_0^{'}$ is rejected if $Y_1^n \geq r + 1$ or $Y_1^n \leq l - 1$, otherwise, H_0' is not rejected. The p-value can be computed as follows: $P(Y_1^n \geq 23 \vee Y_1^n \leq 11 | n = 30, P_0^* = 0.58) = 1 - P(Y_1^n \geq 22, n = 30, P_0^* = 0.58) + P(Y_1^n \leq 11, n = 30, P_0^* = 0.58) = 0.0419 < 0.05$. It is concluded that H_0 can be rejected if if $Y_1^n \geq 23$ or $Y_1^n \leq 11$, at the 0.05 significance level.

The minimum value of the NPI lower reproducibility probability of the event $Y_{n+1}^{2n} \geq r+1$ or $Y_{n+1}^{2n} \leq l - 1$ for this two-sided GB take a value less than 0.5. This small probability of future observations at the 'other end' resulting to null hypothesis rejection could be why the minimum for $\underline{RP}(y)$ for y that does not lead to null hypothesis rejection for $y = 12$ and $y = 22$, respectively, is less than 0.5.

In the case of $y = 0$ or $y = 30$, the NPI upper probability for RP is equal to 1,

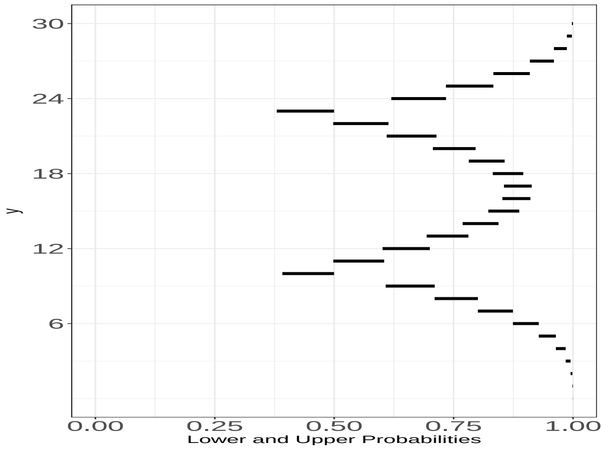


Figure 2.7: NPI-RP of two-sided tests using data collected from the GB method of $n = 30$, $\pi_{A_0} = 0.7, \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.01,\ P_0^* = 0.58$

indicating that such data do not provide evidence against the possibility that there would never be any Yes answers (for $y = 0$), or that there would never be any No answers (for $y = 30$).

Figure 2.6 presents the NPI-RP-GB with the two-sided alternative hypothesis $H_1' : \pi_A \neq 0.7$, with $n = 30$ and at $\alpha = 0.05$. The null hypothesis is rejected if and only if $y \leq 12$ or $y \geq 22$. For the case presented in Figure 2.7, with $n = 30$ and $\alpha = 0.01$, the null hypothesis is rejected if and only if $y \leq 10$ or $y \geq 23$. For values of y for which $H_0^{'}$ is rejected, the NPI lower and upper reproducibility probabilities at significance $\alpha = 0.05$ are smaller than the NPI lower and upper reproducibility probabilities at significance $\alpha = 0.01$, while for values of y for which $H_0^{'}$ is not rejected, they are larger. This is logical because changing the level of significance changes the rejection threshold.

In addition, an increase in γ causes an increase in P_0^* , which leads to an increase in the threshold values, which results in higher lower and upper reproducibility probabilities of the event $Y_{n+1}^{2n} \geq l - 1$ or $Y_{n+1}^{2n} \leq r + 1$.

As a result, when α is small, the null hypothesis is not rejected for a wide range of y-values, which is consistent with the previous discussion of one-sided tests. The NPI lower and upper reproducibility probability are lower for $y-$values for which H_0 is rejected and higher for $y-$values for which H_0 is not rejected, as shown in a comparison of Tables 2.8

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	1.0000	1	10	0.7855	0.8567	21	0.5809	0.6986
1	1.0000	1.0000	11	0.7003	0.7871	22	0.4825	0.6061
2	0.9999	1.0000	12	0.6045	0.7036	23	0.3722	0.4964
3	0.9995	0.9999	13	0.3890	0.4959	24	0.6209	0.7375
4	0.9983	0.9995	14	0.4837	0.5935	25	0.7344	0.8346
5	0.9949	0.9983	15	0.5722	0.6808	26	0.8334	0.9101
6	0.9875	0.9949	16	0.6455	0.7507	27	0.9097	0.9603
7	0.9731	0.9875	17	0.6955	0.7976	28	0.9601	0.9872
8	0.9483	0.9731	18	0.7159	0.8172	29	0.9872	0.9977
9	0.9101	0.9484	19	0.7032	0.8072	30	0.9977	1
10	0.8560	0.9104	20	0.6570	0.7671			

Table 2.8: NPI-RP-FM of $\alpha = 0.05,\ \pi_{A_0} = 0.7,\ \gamma_1 = 0.15,\ \gamma_2 = 0.10,\ l = 13,\ r = 23.$

and 2.9. This makes sense because a change in the rejection threshold logically follows a change in the level of significance α.

Example 2.4.2 This example introduces the reproducibility of two-sided hypothesis tests using data collected from the forced method. Assume that the probability of being asked the sensitive question is 0.75, the forced 'Yes' answer has probability $\gamma_1 = 0.10$ and the forced 'No' answer has probability $\gamma_2 = 0.15$.

Assume that we have a sample of size $n = 30$, and we want to test the null hypothesis that the proportion of people who have the sensitive characteristic $\pi_A = 0.7$, against the alternative hypothesis that $\pi_A \neq 0.7$.

For $\alpha = 0.05$ or $\alpha = 0.01$, we consider NPI-RP for the two-sided hypothesis tests using data collected from the FM method as follows.

The null and alternative hypotheses of the proportion π_A of interest are:

$$H_0' : \pi_A = 0.7 \quad \text{and} \quad H_1' : \pi_A \neq 0.7 \tag{2.28}$$

The corresponding null and alternative hypotheses for the proportion of 'Yes' answers, using Equation (1.8), are:

$$H_0 : P^* = 0.625 \quad \text{and} \quad H_1 : P^* \neq 0.625 \tag{2.29}$$

y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$	y	$\underline{RP}(y)$	$\overline{RP}(y)$
0	0.9999	1	11	0.3943	0.4999	22	0.7295	0.8179
1	0.9995	0.9999	12	0.4994	0.6035	23	0.6241	0.7332
2	0.9976	0.9995	13	0.6024	0.6989	24	0.4993	0.6257
3	0.9923	0.9976	14	0.6965	0.7812	25	0.3651	0.4999
4	0.9805	0.9923	15	0.7762	0.8473	26	0.6347	0.7642
5	0.9580	0.9805	16	0.8378	0.8960	27	0.7642	0.8729
6	0.9210	0.9580	17	0.8788	0.9272	28	0.8729	0.9486
7	0.8666	0.9210	18	0.8978	0.9416	29	0.9486	0.9881
8	0.7941	0.8666	19	0.8935	0.9390	30	0.9881	1
9	0.7056	0.7942	20	0.8648	0.9189			
10	0.6054	0.7056	21	0.8103	0.8793			

Table 2.9: NPI-RP-FM of $\alpha = 0.01$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.15$, $\gamma_2 = 0.10$, $l = 11$, $r = 25$.

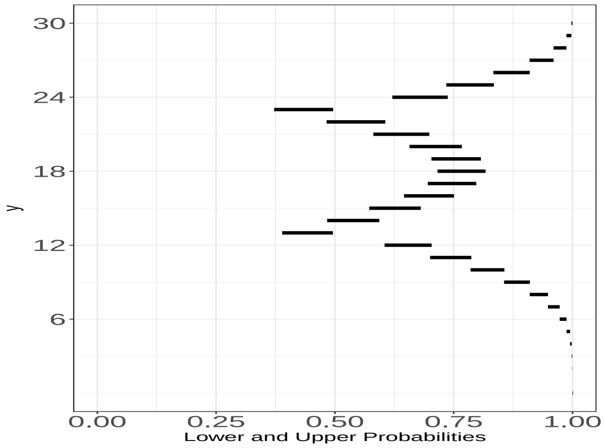


Figure 2.8: NPI-RP of two-sided tests using data collected from the FM method of $n = 30$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.10$, $\gamma_2 = 0.15$, $\alpha = 0.05$, $P_0^* = 0.625$

The p-value can be computed as follows: $P(Y_1^n \geq 24 \vee Y_1^n \leq 12 | n = 30, P_0^* = 0.625) = 1 - P(Y_1^n \geq 23, n = 30, P_0^* = 0.625) + P(Y_1^n \leq 12, n = 30, P_0^* = 0.625) = 0.0428 < 0.05$. It is concluded that H_0 can be rejected if $Y_1^n \geq 24$ or $Y_1^n \leq 12$, at the 0.05 significance level.

The null hypothesis is rejected in Table 2.8 with $n = 30$ and $\alpha = 0.05$ if and only if $y \leq 12$ or $y \geq 24$ are true. The lower and upper reproducibility probabilities are minimum at these values, however, it is important to note that these lower probabilities are no longer exactly equal to 0.5 as was the case for the one-sided alternative hypothesis as shown in Section 2.4 for the same reasons are discussed in Example 2.4.1. The minimum value for $\underline{RP}(y)$ for y leads to rejection of the null hypothesis $y = 13$ and $y = 23$. For the case when

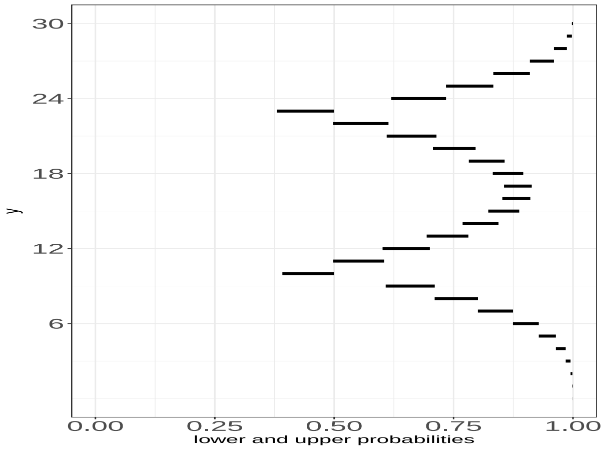


Figure 2.9: NPI-RP of two-sided tests using data collected from the FM method of $n = 30$, $\pi_{A_0} = 0.7$, $\gamma_1 = 0.10$, $\gamma_2 = 0.15$, $\alpha = 0.01$, $P_0^* = 0.625$

the original test did not reject the null hypothesis, the maximum value for $\overline{RP}(y)$ for y such that H_0 is not rejected is 0.8172.

When the significance level changes from $\alpha = 0.05$ to $\alpha = 0.01$, the rejection threshold values l and r change from $13,\ 23$ to $11,\ 25$. Therefore, with small, the null hypothesis is not rejected for a large range of values of l and r, which is consistent with the same property as previously mentioned for one-sided tests. For values of y for which H_0 is rejected, and for values of y for which H_0 is not rejected, the NPI lower and upper reproducibility probabilities are smaller and larger, respectively. This makes sense because a change in the rejection threshold logically follows a change in the level of significance as Tables 2.8 and 2.9 and Figures 2.8 and 2.9 are shown.

2.5 A measure of reproducibility for statistical hypothesis tests

One objective of the reproducibility of hypothesis tests based on RRT methods is to compare RRT methods. This is non-trivial particularly if the different RRT methods require different sample sizes to achieve a similar level of significance and power for a specific alternative hypothesis. We present a new measure of reproducibility for this objective. This measure can be based on either the NPI lower reproducibilities probability or the NPI upper reproducibilities probability, we call it the measure of reproducibility probability (MRP). This measure of reproducibility probability can be applied to one-sided and two-sided hypothesis tests.

2.5.1 A measure of reproducibility for one-sided hypothesis tests

The measure of the lower reproducibility probability under H_0 (MRP$^l_0(z)$) is the probability, under H_0, for the event that $\underline{RP}(Y) \geq z$, at the value $z \in [0,1]$. It is just a probability at a particular level of z. If we assume that the MRP$^l_0(z)$ is an appropriate measurement in this situation, we gather all reproducibility probabilities at z.

Therefore, with a sample with size n and probability of 'Yes' answer P_0^* under H_0, MRP$^l_0$ under H_0 for one-sided test is

$$\begin{aligned} MRP^l_0(z) =& P(\underline{RP}(Y) \geq z | H_0) = P[\underline{RP}(Y) > z \; | Y \sim \text{Bin}(n, P_0^*)] \\ =& 1 - \sum_{y=a(z)}^{b(z)} \binom{n}{y} (P_0^*)^y (1-P_0^*)^{n-y} \end{aligned} \tag{2.30}$$

for $z \in [0,1]$ and P_0^* is derived from Equation (1.5) and (1.8) in Section 1.2.1 under H_0. We specify all the values of $Y = y$ for which $\underline{RP}(Y) \geq z$ by removing all the values in the two intervals $[0, 1, 2, ..., a(z)-1]$, $[b(z)+1, ..., n]$ which are not included in the interval $[a(z), b(z)]$. We assume that $a(z)$ is integer such that $\underline{RP}(y) \geq z$ for $y < a(z)$ and $\underline{RP}(a(z)) < z$, and $b(z)$ is also integer such that $\underline{RP}(y) \geq z$ for $y > b(z)$ and $\underline{RP}(b(z)) < z$.

Similarly, the measure of the upper reproducibility probability under H_0 (MRP$^u_0(z)$) under H_0 for one-sided test is

$$\begin{aligned} MRP^u_0(z) =& P(\overline{RP}(Y) \geq z | H_0) = P[\overline{RP}(Y) > z \; | Y \sim \text{Bin}(n, P_0^*)] \\ =& 1 - \sum_{y=a(z)}^{b(z)} \binom{n}{y} (P_0^*)^y (1-P_0^*)^{n-y} \end{aligned} \tag{2.31}$$

We specify all the value of $Y = y$ for which $\overline{RP}(Y) \geq z$ by removing all the values in the two intervals $[0, 1, 2, ..., a(z) - 1]$, $[b(z) + 1, ..., n]$ which are not included in the interval $[a(z), b(z)]$, where $a(z)$ is integer such that $\overline{RP}(y) \geq z$ for $y < a(z)$ and $\overline{RP}(a(z)) < z$, and $b(z)$ is also integer such that $\overline{RP}(y) \geq z$ for $y > b(z)$ and $\overline{RP}(b(z)) < z$.

So in the last explanation, we investigates the measure of reproducibility under the null hypothesis, H_0 which is a statistical proposition stating that there is no significant difference between P^* and P^*_0. Now, we need to investigate the measure of reproducibility under the alternative hypothesis, H_1, is a statistical proposition stating that there is a significant difference between P^* and P^*_0 that means $P^* > P^*_0$. Similarly, to compute the measure of lower reproducibility probability under H_1, we use Equations (2.32) and (2.33):

$$\begin{aligned} MRP^l_1(z) =& P(\underline{RP}(y) \geq z|H_1) = P[\underline{RP}(y) > z\ |Y \sim Bin(n, P^*_1)] \\ =& 1 - \sum_{y=a(z)}^{b(z)} \binom{n}{y} (P^*_1)^y (1 - P^*_1)^{n-y} \end{aligned} \tag{2.32}$$

Similarly, the measure of the upper reproducibility probability under H_0 (MRP$^u_1(z)$) under H_1 for one-sided test is

$$\begin{aligned} MRP^u_1(z) =& P(\overline{RP}(Y) \geq z|H_1) = P[\overline{RP}(Y) > z\ |Y \sim Bin(n, P^*_1)] \\ =& 1 - \sum_{y=a(z)}^{b(z)} \binom{n}{y} (P^*_1)^y (1 - P^*_1)^{n-y} \end{aligned} \tag{2.33}$$

where P^*_1 is derived from Equations (1.5) and (1.8) in Section 1.2.1 under H_1 in which power is computed. Examples 2.5.1 and 2.5.2 illustrate this measurement using the GB and the FM methods as explained in Section 1.2.1.

Example 2.5.1 This example illustrates the measure of reproducibility probability for one-sided hypothesis tests using data collected from the GB method [2]. With a sample of size $n = 30$, assume that we interested in a sensitive characteristic A, the probability of a person having the sensitive characteristic is π_A, and the known proportion of those who have the unrelated characteristic is $\pi_B = 0.30$. Suppose that we use a randomising device with a probability of $\gamma = 0.7$ for the sensitive question being asked.

z	$\text{MRP}_0^l(z)$	z	$\text{MRP}_0^l(z)$	z	$\text{MRP}_0^l(z)$
0.5000	0.9020	0.8954	0.3667	0.9939	0.0151
0.6106	0.8067	0.9101	0.2400	0.9956	0.0056
0.6145	0.7898	0.9449	0.1420	0.9980	0.0018
0.7102	0.6644	0.9479	0.1419	0.9990	0.0018
0.7240	0.6575	0.9680	0.0755	0.9992	0.0005
0.7941	0.5137	0.9790	0.0754	0.9997	0.0001
0.8198	0.5115	0.9824	0.0358	0.9999	0.0000
0.8605	0.3673	0.9909	0.0151	1.0000	0.0000

Table 2.10: $\text{MRP}_0^l(z)$ with GB data of $n = 30,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,$ $\alpha = 0.05,\ P_0^* = 0.7, P_1^* = 0.72$

z	$\text{MRP}_1^l(z)$	z	$\text{MRP}_1^l(z)$	z	$\text{MRP}_1^l(z)$
0.5000	0.6866	0.8605	0.0721	0.9824	0.0009
0.6106	0.5618	0.8954	0.0392	0.9909	0.0007
0.6145	0.4181	0.9101	0.0254	0.9939	0.0001
0.7102	0.3299	0.9449	0.0197	0.9956	0.0001
0.7240	0.2221	0.9479	0.0071	0.9980	0.0001
0.7941	0.1678	0.9680	0.0050	0.9990	0.0000
0.8198	0.1013	0.9790	0.0016		

Table 2.11: $\text{MRP}_1^l(z)$ with GB data of $n = 30,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,$ $\alpha = 0.05,\ P_0^* = 0.7, P_1^* = 0.72$

We want to test the null and alternative hypotheses :

$$H_0' : \pi_A = 0.7 \ \text{ and } \ H_1' : \pi_A > 0.7 \tag{2.34}$$

The corresponding null and alternative hypotheses for P^* using Equation (1.5) for NPI-RP-GB are:

$$H_0 : P^* = 0.58 \ \text{ and } \ H_1 : P^* > 0.58 \tag{2.35}$$

We takes a specific value under the null hypothesis $H_0^{'}$ which is $\pi_{A_0} = 0.7$ and we takes a specific value under the alternative hypothesis $H_1^{'}$ which is that $\pi_{A_1} = 0.9$ where the proportion of people saying 'Yes' are

$$\begin{aligned} P_0^* &= \gamma\pi_{A_0} + (1-\gamma)\pi_B = 0.58 \\ P_1^* &= \gamma\pi_{A_1} + (1-\gamma)\pi_B = 0.72 \end{aligned} \tag{2.36}$$

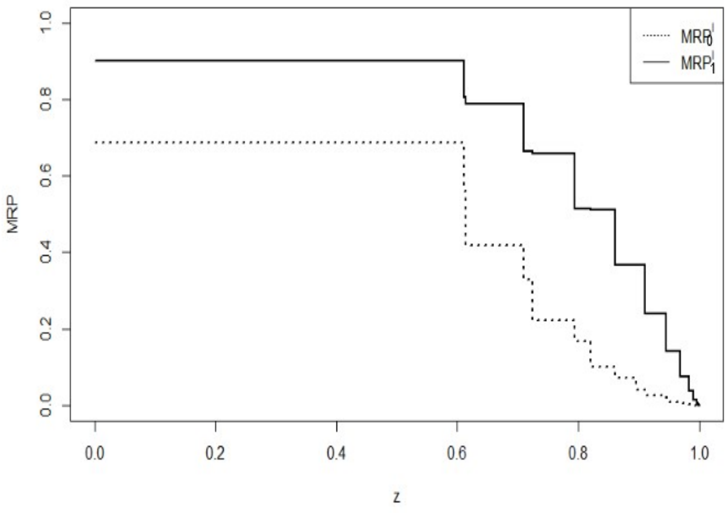


Figure 2.10: $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ with GB data of $n = 30$, $\pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,$ $\pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P_0^* = 0.7, P_1^* = 0.72$

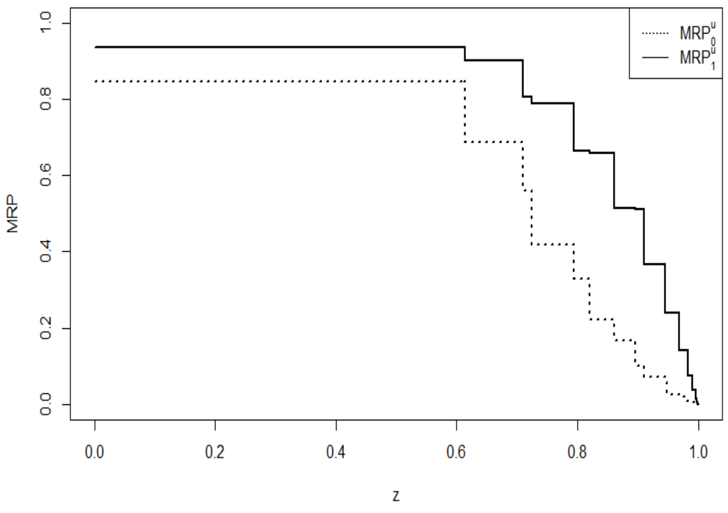


Figure 2.11: $\mathrm{MRP}_0^u(z)$ and $\mathrm{MRP}_1^u(z)$ with GB data of $n = 30$, $\pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,$ $\pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P_0^* = 0.58, P_1^* = 0.72$

Then, we calculate the $\underline{RP}(Y)$ and $\overline{RP}(Y)$. Then, we compute $\mathrm{MRP}_0^l(z)$ by calculating the probability of all $\underline{RP}(Y) \geq z$ with $Y \sim \mathrm{Bin}(n, \pi_{A_0})$. Similarly, we compute $\mathrm{MRP}_1^l(z)$ by calculating the probability of all $\underline{RP}(Y) \geq z$ with $Y \sim \mathrm{Bin}(n, \pi_{A_1})$ using Equations (2.30) and (2.32) respectively.

Figure 2.10 and Tables 2.10 and 2.11 reveal MRP_0^l and MRP_1^l for this test scenario. It has been noted that $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ decrease if z increases. $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ have higher value for z between 0 and 0.6. Both $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ get values close to 0 at $z = 1$. MRP_0^l get values higher than MRP_1^l for $z \in [0, 1]$. Changes in the $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ are caused by further variations of γ, π_{A_0}, π_B and α. Increasing their values which leads to increasing the $\text{MRP}_1^l(z)$ then which make $\text{MRP}_1^l(z)$ close to $\text{MRP}_0^l(z)$.

Similarly, $\text{MRP}_0^u(z)$ and $\text{MRP}_1^u(z)$ of the GB can be derived as the same as $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ derivation. It can be noted that $\text{MRP}_0^u(z)$ and $\text{MRP}_u^l(z)$ get higher values than $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ because the probability of yes-response in the FM method is larger than the probability of yes-response in the GB method, the threshold value gets larger and consequently the non-rejection region and the reproducibility are increased as shown in Figures 2.10 and 2.11.

Example 2.5.2 This example illustrates the measure of reproducibility probability of one-sided hypothesis tests using data collected from the FM method for a sample size $n = 30$. Assume that the probability of being asked the sensitive question is $\gamma = 0.75$, the forced 'Yes' answer has probability $\gamma_1 = 0.10$ and the forced 'No' answer has probability $\gamma_2 = 0.15$.

We assume the null and the alternative hypotheses as follows:

$$H_0' : \pi_A = 0.7 \quad \text{and} \quad H_1' : \pi_A > 0.7 \tag{2.37}$$

The corresponding null and alternative hypotheses for P^* using Equation (1.8) for NPI-RP-FM are:

$$H_0 : P^* = 0.625 \quad \text{and} \quad H_1 : P^* > 0.625 \tag{2.38}$$

We takes a specific value under the null hypothesis $H_0^{'}$ which is $\pi_{A_0} = 0.7$ and we takes a specific value under the alternative hypothesis $H_1^{'}$ which is that $\pi_{A_1} = 0.9$ where the

z	$\text{MRP}^l_0(z)$	z	$\text{MRP}^l_0(z)$	z	$\text{MRP}^l_0(z)$
0.5000	0.8832	0.9097	0.3148	0.9959	0.0036
0.6145	0.7747	0.9149	0.1971	0.9977	0.0036
0.6195	0.7539	0.9483	0.1114	0.9981	0.0011
0.7163	0.6172	0.9601	0.1112	0.9992	0.0003
0.7340	0.6088	0.9702	0.0564	0.9997	0.0001
0.8007	0.4597	0.9837	0.0255	0.9999	0.0000
0.8333	0.4571	0.9872	0.0255	1.0000	0.0000
0.8666	0.3154	0.9916	0.0102		

Table 2.12: $\text{MRP}^l_0(z)$ with FM data of $n = 30$, $\gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,$ $\alpha = 0.05,\ P^*_0 = 0.625,\ P^*_1 = 0.775$

z	$\text{MRP}^l_1(z)$	z	$\text{MRP}^l_1(z)$	z	$\text{MRP}^l_1(z)$
0.5000	0.6899	0.8333	0.0933	0.9702	0.0049
0.6145	0.5898	0.8666	0.0785	0.9837	0.0047
0.6195	0.4200	0.9097	0.0311	0.9872	0.0005
0.7163	0.3589	0.9149	0.0251	0.9916	0.0005
0.7340	0.2185	0.9483	0.0230	0.9959	0.0005
0.8007	0.1863	0.9601	0.0055	0.9977	0.0000

Table 2.13: $\text{MRP}^l_0(z)$ with FM data of $n = 30$, $\gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,$ $\alpha = 0.05,\ P^*_0 = 0.625,\ P^*_1 = 0.775$

proportion of people saying ‘Yes’ are

$$\begin{aligned} P^*_0 &= \gamma_1 + (1 - \gamma_1 - \gamma_2)\pi_{A_0} = 0.625 \\ P^*_1 &= \gamma_1 + (1 - \gamma_1 - \gamma_2)\pi_{A_1} = 0.775 \end{aligned} \tag{2.39}$$

Then, we calculate the $\underline{RP}(Y)$ and $\overline{RP}(Y)$. Then, we compute $\text{MRP}^l_0(z)$ by calculating the probability of all $\underline{RP}(Y) \geq z$ with $Y \sim Bin(n, \pi_{A_0})$. Similarly, e compute $\text{MRP}^l_1(z)$ by calculating the probability of all $\underline{RP}(Y) \geq z$ with $Y \sim Bin(n, \pi_{A_1})$ using Equations (2.30) and (2.32) respectively.

Figure 2.12 and Tables 2.12 and 2.13 reveal MRP^l_0 and MRP^l_1 for this test scenario. It has been noted that $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ decrease if z increases. $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ have higher value for z between 0 and 0.6. Both $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ get values close to

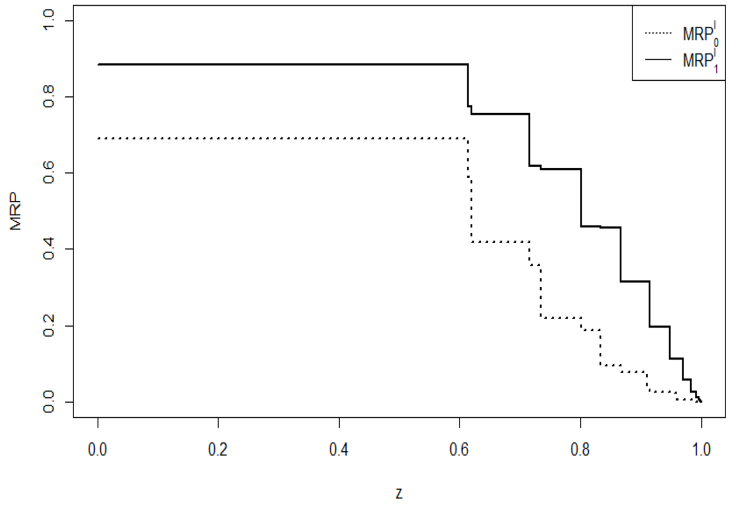


Figure 2.12: $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ with FM of $n = 30$, $\gamma_2 = 0.10$, $\gamma_1 = 0.15$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $P^*_0 = 0.625$, $P^*_1 = 0.775$

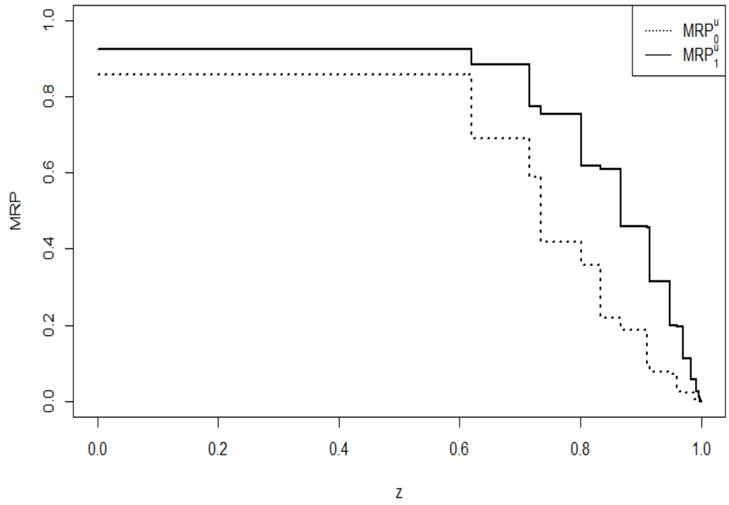


Figure 2.13: $\text{MRP}^u_0(z)$ and $\text{MRP}^u_1(z)$ with FM of $n = 30$, \$$\gamma_2 = 0.10$, $\gamma_1 = 0.15, \pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $P_0 = 0.625$, $P_1 = 0.775$

0 at $z = 1$. MRP^l_0 get values higher than MRP^l_1 for $z \in [0, 1]$. Changes in the $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ are caused by further variations of γ_1, γ_2, π_{A_0} and α. Increasing their values leads to increasing $\text{MRP}^l_1(z)$ then that makes $\text{MRP}^l_1(z)$ close to $\text{MRP}^l_0(z)$.

From the results of $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the GB and the FM method, it is noted that the $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the GB method is higher than $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the FM method as shown in Tables 2.10, 2.11, 2.12 and 2.13.

Similarly, $\mathrm{MRP}_0^u(z)$ and $\mathrm{MRP}_1^u(z)$ of the FM method can be derived as the same as $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ derivation. It can be noted that $\mathrm{MRP}_0^u(z)$ and $\mathrm{MRP}_u^l(z)$ get higher values than $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ as shown in Figures 2.12 and 2.13.

For comparison between the GB and FM method using the same sample size, we can noticed that $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the FM method is greater than $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the GB method. However, this measurement needs to apply using the required sample size with high power and at a specific significance level.

when comparing the GB and the FM methods with the same sample size, we find that the $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of the FM technique are higher than those of the GB method. However, in order for this measurement to be meaningful, the minimum required sample size must be used, along with high power and a certain level of significance. as explained in Section 2.5.2.

2.5.2 The minimum required sample size of measurement of reproducibility probability

In the last explanation, we derive the NPI lower and the upper reproducibility probabilities and define the measurement of $\mathrm{MRP}(z)$. Now, we need to select the best parameters of the RRT methods, ones that result in tests with appropriate power and p-values to determine the required minimum sample size, which leads to increasing reproducibility probability of hypothesis tests.

Assume that Y is the random quantity where $Y \sim \mathrm{Bin}(n_k, P_0^*)$, and the null hypothesis is $H_0 : P^* = P_0^*$ versus the alternative hypothesis is $H_1 : P^* > P_0^*$. The required sample size n will be large. So, we use the normal approximation instead of the binomial distribution when $nP_0^* \geq 5$ and $n(1 - P_0^*) \geq 5$ are both true. We standardise the random quantity Y as $Z = \frac{Y - E(Y)}{\sqrt{\mathrm{Var}(Y)}}$. Then, for the approximate one-sided test of H_0, we reject H_0 when the p-value of the test statistic $\dot{z}$ is less than or equal α where $\dot{z} = \frac{\hat{P}-P_0^*}{\sqrt{\frac{P_0^*(1-P_0^*)}{n}}}$ and $\hat{P}$ is the sample proportion.

Suppose H_1 is true with $P^* = P_1$. Then, the approximate power is calculated using [21]:

$$1-\beta \approx P\Bigg(Z \geq \frac{n(P_0^*-P_1^*)+z_{1-\alpha}\sqrt{nP_0^*(1-P_0^*)}}{\sqrt{nP_1^*(1-P_1^*)}}\Bigg) \tag{2.40}$$

where the values of $z_{1-\alpha}$ indicates to the $(1-\alpha) \times 100$ percentiles of standard normal distribution.

Now, we determine the minimum sample size n_k required for this case for getting an approximate power (i.e. $1-\beta$) at a level of significance of (i.e., $\alpha = 0.05$) using Equation (2.41) [21].

$$\lceil n_k \rceil \geq \left[\frac{z_{1-\alpha}\sqrt{P_0^*(1-P_0^*)}+z_{1-\beta}\sqrt{P_1^*(1-P_1^*)}}{P_1^*-P_0^*}\right]^2 \tag{2.41}$$

where $\lceil n_k \rceil$ is the smallest integer greater than or equal to n_k, $P_0^*,\ P_1^*$ are the proportion of people who have the sensitive characteristics under H_0 and H_1 respectively. The values of $z_{1-\alpha}$ and $z_{1-\beta}$ indicate to the $(1-\alpha) \times 100$ and $(1-\beta) \times 100$ percentiles of standard normal distribution respectively. If the hypothesis tests do not give the required power equal to or greater than 0.90 using sample size n_k, Fleiss, Levin, and Paik [54] suggested adding $\frac{1}{|P_1^*-P_0^*|}$ as a continuity correction to $\lceil n_k \rceil$ to get the required power as follows.

$$n = \lceil n_k \rceil + \frac{1}{|P_1^*-P_0^*|} \tag{2.42}$$

Using the threshold value C in Equation (2.11), we calculate p-value as follows:

$$P(Y_1^n \geq C \mid P^* = P_0^*) = 1 - \sum_{y=C}^{n} \binom{n}{y} (P_0^*)^y (1-P_0^*)^{n-y} \tag{2.43}$$

we can use the exact power of $P^* = P_1^*$ [54] as follows:

$$P(Y_1^n \geq C \mid P^* = P_1^*) = \sum_{y=C}^{n} \binom{n_k}{y} (P_1^*)^y (1-P_1^*)^{n_k-y} \tag{2.44}$$

Similarly, for two-sided tests, assume that the null hypothesis is $H_0 : P^* = P_0^*$ versus the alternative hypothesis is $H_1 : P^* \neq P_0^*$. We use the normal approximation instead of the binomial distribution when $nP_0^* \geq 5$ and $n(1-P_0^*) \geq 5$ are both true.

The approximate power is

$$1-\beta \approx \Phi(\dot{z}-z_{1-\frac{\alpha}{2}})+\Phi(-\dot{z}-z_{1-\frac{\alpha}{2}}) \tag{2.45}$$

where Φ is the standard normal distribution function and $\dot{z}$ is the test statistics.

Now, we determine the minimum sample size n_k required for this case for getting an approximate power (i.e. $1-\beta$) at a level of significance of (i.e., $\alpha = 0.05$) as follows:

$$\lceil n_k \rceil \geq (1-P_1^*) \times P_1^* \left[\left(\frac{z_{1-\frac{\alpha}{2}} + z_{1-\beta}}{P_1^* - P_0^*} \right) \right]^2 \tag{2.46}$$

where the values of $z_{1-\frac{\alpha}{2}}$ and $z_{1-\beta}$ indicates to the $(1-\frac{\alpha}{2}) \times 100$ and $(1-\beta) \times 100$ percentiles of standard normal distribution respectively.

Using the threshold value in Equations (2.20) and (2.21), the p-value of test for the proportion $P^* = P_0^*$ is

$$\text{p-value} = 2P(Y_1^n \geq r \mid P^* = P_0^*) = 2\sum_{y=r}^{n} \binom{n}{y} (P_0^*)^y (1-P_0^*)^{n-y} \tag{2.47}$$

$$\begin{aligned} P(Y_1^n \leq l-1 \mid P^* = P_1^*) + P(Y_1^n \geq r \mid P^* = P_1)] = \\ = \sum_{y=0}^{l-1} \binom{n}{y} (P_1^*)^y (1-P_1^*)^{n-y} + \sum_{y=r}^{n} \binom{n}{y} (P_1^*)^y (1-P_1^*)^{n-y} \end{aligned} \tag{2.48}$$

Using the same parameters of the GB and FM methods in Examples 2.5.2 and 2.5.1, we calculate the minimum required sample sizes with power equals to 0.90 and p-value is less than 0.05 as Figures 2.14 and 2.15 show. It is noted that we have higher $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ of the GB and the FM method with higher power 0.9356, 0.9417 and p-values 0.0427, 0.0345 respectively. $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ are close to each other over all z.

This measure can be applied of $\text{MRP}_0^u(z)$ and $\text{MRP}_1^u(z)$ for two-sided hypothesis tests based on data collected from RRT methods for the range values of $Y = y$ in which $\underline{RP}(Y) \geq z$ and $\overline{RP}(Y) \geq z$ respectively.

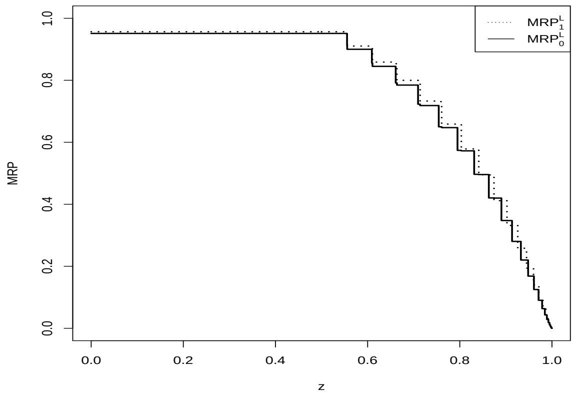


Figure 2.14: $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of one-sided hypothesis tests using the GB method of $n = 113,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P_0^* = 0.58, P_1^* = 0.72$

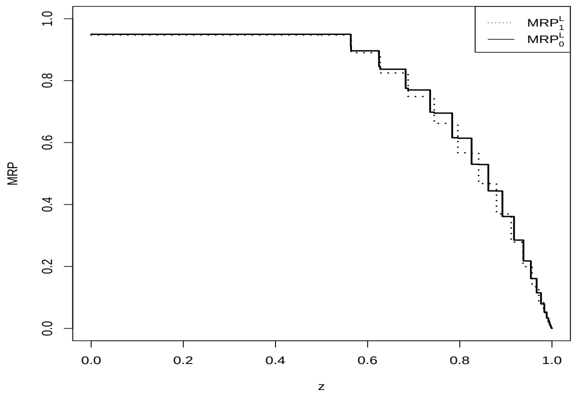


Figure 2.15: MRP_0^l and MRP_1^l of one-sided hypothesis tests using FM method of $n = 93$, $\gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \alpha = 0.05,\ P_0^* = 0.625,\ P_1^* = 0.775$

2.5.3 A measure of reproducibility for two-sided hypothesis tests

This section introduces the measure of reproducibility for two-sided hypothesis tests using data collected from the GB and FM methods using two threshold values $l,\ r$ to calculate $\mathrm{MRP}_0^l(z)$ and MRP_1^l(z) based on NPI lower and upper reproducibility probabilities under H_0 and H_1 respectively. In addition, we calculate $\mathrm{MRP}_0^u(z)$ and $\mathrm{MRP}_1^u(z)$ under H_0 and

z	$\mathrm{MRP}_0^l(z)$	z	$\mathrm{MRP}_0^l(z)$	z	$\mathrm{MRP}_0^l(z)$
0.3764	0.9830	0.6428	0.5266	0.8200	0.0020
0.3872	0.9434	0.6501	0.4314	0.8607	0.0007
0.4816	0.9080	0.6912	0.2873	0.8954	0.0005
0.4824	0.8415	0.6960	0.1619	0.9149	0.0001
0.5706	0.7436	0.7029	0.1524	0.9479	0.0001
0.5761	0.6810	0.7099	0.0086	0.9527	0.0000
0.6058	0.6602	0.7246	0.0063		
0.6162	0.6534	0.7893	0.0026		

Table 2.14: $\mathrm{MRP}_0^l(z)$ of two-sided hypothesis tests using data collected from the GB method of $n = 30$, $\pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P_0^* = 0.7, P_1^* = 0.72$

H_1 for all z.

We use the mentiond Equations (2.30) and (2.32) to calculate the $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ and Equations (2.31) and (2.33) to calculate the $\mathrm{MRP}_0^u(z)$ and $\mathrm{MRP}_1^u(z)$ under H_1 using three intervals $[0, l-1]$, the middle region between the threshold values $[l, r]$ and $[r+1, n]$ where l and r are calculated from Equation 2.20 and 2.21.

Example 2.5.3 This example illustrates the measure of reproducibility probability for two-sided hypothesis tests using data collected from the GB method [2]. With a sample of size $n = 30$, assume that we are interested in a sensitive characteristic A, the probability of a person having the sensitive characteristic is π_A, and the known proportion of those who have the unrelated characteristic is $\pi_B = 0.30$. Suppose that we use a randomising device with a probability of $\gamma = 0.7$ for the sensitive question being asked.

Tables 2.14 and 2.15 show the $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ using data collected from the GB method. It is noted that $\mathrm{MRP}_0^l(z)$ takes the maximum value 0.9830 of $z = 0.3764$ whereas $\mathrm{MRP}_1^l(z)$ takes the maximum value 0.8563 of $z = 0.3764$, then they decrease till 0.0000 for $z = 0.9527$ of $\mathrm{MRP}_0^l(z)$ and value $z = 0.9939$ for $\mathrm{MRP}_1^l(z)$.

For sample size n, it is noticed that $\mathrm{MRP}_0^l(z) = 0.6810$ for $z = 0.5761$ which is close to $\mathrm{MRP}_1^l(z) = 0.6871$ for $z = 0.5706$. In addition, $\mathrm{MRP}_0^l(z) = 0.2873$ for $z = 0.6912$ which is close to $\mathrm{MRP}_1^l(z) = 0.2879$ for $z = 0.6501$. Further points between $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$

z	$\mathrm{MRP}^l_0(z)$	z	$\mathrm{MRP}^l_0(z)$	z	$\mathrm{MRP}^l_0(z)$
0.3764	0.8563	0.6501	0.2879	0.8954	0.0041
0.3872	0.8556	0.6912	0.2586	0.9149	0.0041
0.4816	0.6949	0.6960	0.1704	0.9479	0.0007
0.4824	0.6929	0.7029	0.1704	0.9527	0.0007
0.5706	0.6871	0.7099	0.1161	0.9765	0.0007
0.5761	0.5344	0.7246	0.0496	0.9791	0.0001
0.6058	0.5342	0.7893	0.0495	0.9898	0.0001
0.6162	0.4264	0.8200	0.0166	0.9939	0.0000
0.6428	0.4126	0.8607	0.0166		

Table 2.15: $\mathrm{MRP}^l_1(z)$ of two-sided hypothesis tests using data collected from the GB method of $n = 30$, $\pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P^*_0 = 0.7, P^*_1 = 0.72$

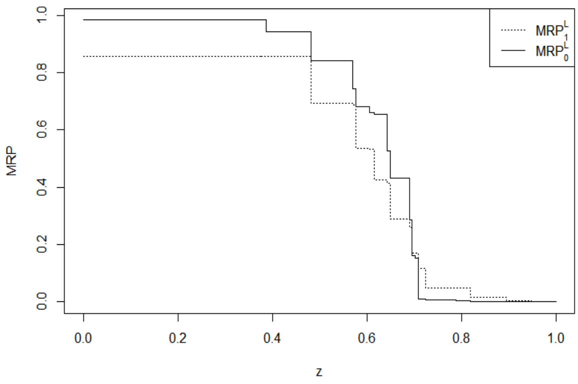


Figure 2.16: $\mathrm{MRP}^l_0(z)$ and $\mathrm{MRP}^l_1(z)$ of two-sided hypothesis tests using data collected from the GB method of $n = 30$, $\pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P^*_0 = 0.7, P^*_1 = 0.72$

are close to each other as shown in Figure 2.16. In almost $\mathrm{MRP}^l_0(z)$ is larger than $\mathrm{MRP}^l_1(z)$ for each value of z.

Conversely, $\mathrm{MRP}^l_0(z)$ is less than $\mathrm{MRP}^l_1(z)$ for each value of z using the required minimum sample size $n = 131$ as shown in Figure 2.17.

Example 2.5.4 This example illustrates the measure of reproducibility probability of two-sided hypothesis tests using data collected from the FM method for a sample size $n = 30$. Assume that the probability of being asked the sensitive question is $\gamma = 0.75$, the

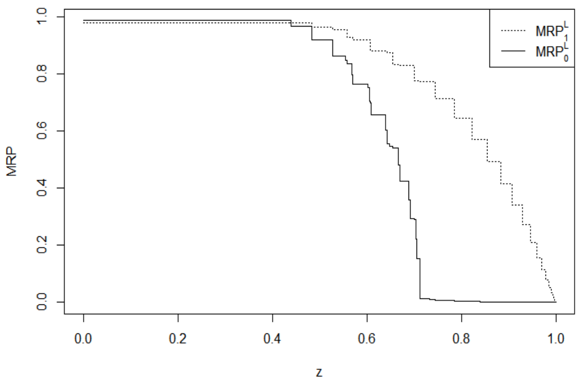


Figure 2.17: $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ of two-sided hypothesis tests using data collected from the GB method of $n = 131,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \pi_B = 0.3,\ \gamma = 0.7,\ \alpha = 0.05,\ P^*_0 = 0.7, P^*_1 = 0.72$

z	$\text{MRP}^l_0(z)$	z	$\text{MRP}^l_0(z)$	z	$\text{MRP}^l_0(z)$
0.3722	0.9792	0.6209	0.6673	0.7344	0.0044
0.3890	0.9483	0.6455	0.5496	0.7855	0.0019
0.4825	0.9054	0.6570	0.4411	0.8334	0.0013
0.4837	0.8506	0.6955	0.2995	0.8560	0.0004
0.5722	0.7648	0.7003	0.2929	0.9097	0.0003
0.5809	0.6909	0.7032	0.1562	0.9101	0.0001
0.6045	0.6756	0.7159	0.0071	0.9483	0.0000

Table 2.16: $\text{MRP}^l_0(z)$ of two-sided hypothesis tests using data collected from the FM method of $n = 30,\ \gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \alpha = 0.05,\ P^*_0 = 0.625,\ P^*_1 = 0.775$

forced 'Yes' answer has probability $\gamma_1 = 0.10$ and the forced 'No' answer has probability $\gamma_2 = 0.15$.

Tables 2.16 and 2.17 show the $\text{MRP}^l_0(z)$ and $\text{MRP}^l_1(z)$ using data collected from the FM method. It is noted that $\text{MRP}^l_0(z)$ takes the maximum value 0.9792 of $z = 0.3722$ whereas $\text{MRP}^l_1(z)$ takes the maximum value 0.8302 of $z = 0.3722$, then they decrease till 0.0000 for $z = 0.9483$ of $\text{MRP}^l_0(z)$ and value $z = 0.9483$ for $\text{MRP}^l_1(z)$.

For sample size n, it is noticed that $\text{MRP}^l_0(z) = 0.6673$ for $z = 0.6209$ which is close to $\text{MRP}^l_1(z) = 0.6610$ for $z = 0.4837$. In addition, $\text{MRP}^l_0(z) = 0.0044$ for $z = 0.7344$ which is

z	$\text{MRP}_1^l(z)$	z	$\text{MRP}_1^l(z)$	z	$\text{MRP}_1^l(z)$
0.3722	0.8302	0.6455	0.3708	0.8334	0.0221
0.3890	0.8300	0.6570	0.2707	0.8560	0.0221
0.4825	0.6610	0.6955	0.2559	0.9097	0.0046
0.4837	0.6603	0.7003	0.2558	0.9101	0.0046
0.5722	0.6582	0.7032	0.1948	0.9483	0.0046
0.5809	0.5172	0.7159	0.1626	0.9601	0.0005
0.6045	0.5171	0.7344	0.0696	0.9731	0.0005
0.6209	0.3767	0.7855	0.0696	0.9872	0.0000

Table 2.17: $\text{MRP}_0^l(z)$ of two-sided hypothesis tests using data collected from the FM method of $n = 30,\ \gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \alpha = 0.05,\ P_0^* = 0.625,\ P_1^* = 0.775$

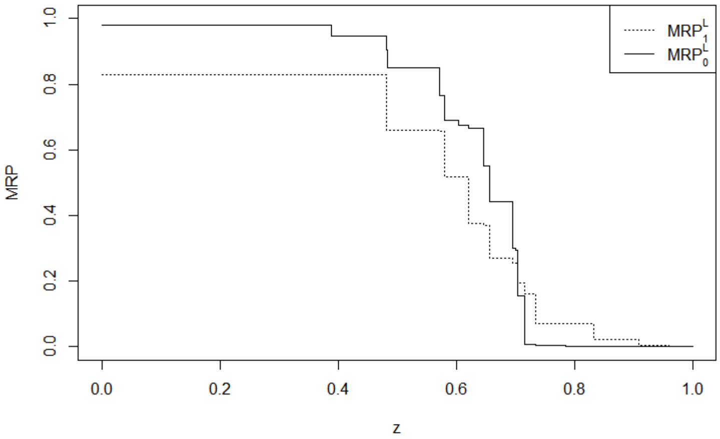


Figure 2.18: $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ of two-sided hypothesis tests using FM data of $n = 30$, $\gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \alpha = 0.05,\ P_0^* = 0.625,\ P_1^* = 0.775$

close to $\text{MRP}_1^l(z) = 0.0046$ for $z = 0.9097$ and $z = 0.9101$. Further points between $\text{MRP}_0^l(z)$ and $\text{MRP}_1^l(z)$ are close for each other as shown in Figure 2.18. In almost $\text{MRP}_0^l(z)$ is larger than $\text{MRP}_1^l(z)$ for each value of z. Conversely, $\text{MRP}_0^l(z)$ is less than $\text{MRP}_1^l(z)$ for each value of z using the required minimum sample size $n = 110$ as shown in Figure 2.19. It is noticed that $\text{MRP}_0^l(z)$ closes from 0 for $z > 0.7$, that happens because the difference between the $\underline{RP}(Y)$ values decrease which provides small measures of $\text{MRP}_0^l(z)$.

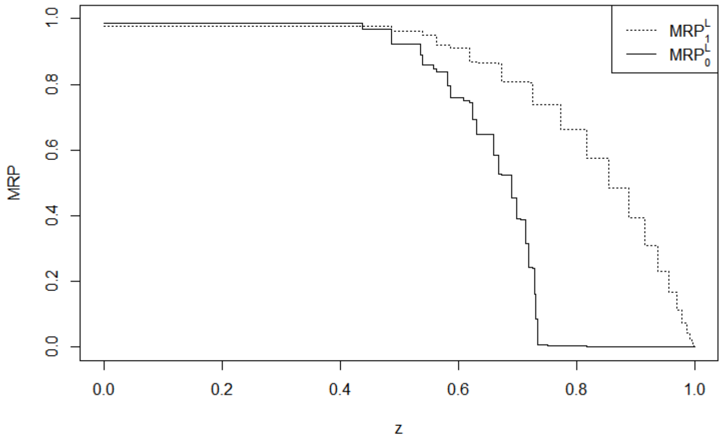


Figure 2.19: $\mathrm{MRP}_0^l(z)$ and $\mathrm{MRP}_1^l(z)$ of two-sided hypothesis tests using FM data of $n = 110$, $\gamma_2 = 0.10,\ \gamma_1 = 0.15,\ \pi_{A_0} = 0.7,\ \pi_{A_1} = 0.9,\ \alpha = 0.05,\ P_0^* = 0.625,\ P_1^* = 0.775$

2.6 Area under MRP of statistical tests based on RRT data

In this section, we introduce the measure of reproducibility for hypothesis tests to allow comparing of different RRT methods by collecting data for such tests we need summary statistics of MRP(z) over all z. We calculate to use the area under MRP over different values of z to compare between of RP of statistical tests based on data collected from RRT method under H_0 and H_1 which are denoted by AUMRP.

To explain this method to measure the whole area, assume that the initial combinations of variables within RRT method such as $\pi_{A_0},\ \pi_{A_1},\ \gamma,\ \pi_B$ of the GB method or $\pi_{A_0},\ \pi_{A_1}, \gamma_1, \gamma_2$ of the FM method.

We can now compute $\mathrm{MPR}_0^l(z)$ under the null hypothesis and $\mathrm{MPR}_1^l(z)$ under the alternative hypothesis using Equations (2.30) and (2.32). The AUMRP_0 and AUMRP_1 are then calculated as follows. Assume that D is the partition of $z \in [0, 1]$, where z_i is real number of the bound of each partition in the number line and $i = 0,\ 1, ..., n$.

$D = \{[z_0, z_1], [z_1, z_2], \ldots, [z_{n-1}, z_n]\}$, and $0 = z_0 < z_1 < z_2 < \cdots < z_n = 1$.

Therefore, AUMRP_0^l and AUMRP_1^l over $[0, 1]$ with partition D are

$$AUMRP_0^l = \sum_{i=1}^{n} MRP_0^l(z_i^*)\,\Delta z_i \tag{2.49}$$

$$AUMRP_1^l = \sum_{i=1}^{n} MRP_1^l(z_i^*)\,\Delta z_i \tag{2.50}$$

where n is the length of partitions and $\Delta z_i = z_i - z_{i-1}$ and $z_i^* \in [z_{i-1}, z_i]$.

Examples 2.6.1 and 2.6.2 introduce AUMRP_0^l, AUMRP_1^l based on data collected from the GB and FM method.

Example 2.6.1 This example derives AUMPR_0^l and AUMRP_1^l of one-sided hypothesis tests based on data collected from the Greenberg method. Assume that some individuals of a population have a sensitive characteristic A, with π_A as the proportion of the sensitive characteristic in a population whereas π_B is the proportion of unrelated characteristic. We use a randomisation device with a probability of $\gamma = 0.7$ for the sensitive question.

Assume that hypothesised value of the proportion of people with the sensitive characteristic under H_0 and H_1 are $\pi_{A_0} = 0.7$ and $\pi_{A_1} = 0.9$ respectively with the significance level $\alpha = 0.05$ and power 0.90.

Table 2.18 gives the required minimum sample sizes for different values of π_B. At $\pi_B = 0.10$, the threshold value is 71, then the AUMRP_0^l equals 0.8190 and AUMRP_1^l equals 0.8114 with power is 0.9123 and p-value is 0.0383, whereas the AUMRP_0^l equals to 0.8225 and AUMRP_1^l equals to 0.8087 for $\pi_B = 0.25$ with threshold value is 74 and power is 0.9090 and p-value is 0.0356. It is noted that for all values of $\pi_B \in [0, 0.6]$, AUMRP_0^l and AUMRP_1^l taking values between 0.80 and 0.81 and the AUMRP_1^l is always greater than the AUMRP_0^l except the case of $\pi_B = 0.1$.

Example 2.6.2 This example derives AUMPR_0^l and AUMRP_1^l of one-side hypothesis tests using data collected from the FM method. Assume that the probability of being asked the sensitive question is 0.75, the forced ‘Yes’ answer is $\gamma_1 = 0.10$ and the forced ‘No’ answer is $\gamma_2 = 0.15$, where the significance level is $\alpha = 0.05$, and power 0.90.

π_B	**0**	**0.1**	**0.25**	**0.3**	**0.45**	**0.6**
n	121	119	115	113	106	98
C	68	71	74	74	74	73
P_0^*	0.490	0.520	0.565	0.580	0.625	0.670
P_1^*	0.630	0.660	0.705	0.720	0.765	0.810
p-value	0.0469	0.0383	0.0356	0.0427	0.0472	0.0435
power	0.9262	0.9123	0.9090	0.9227	0.9316	0.9313
AUMRP$_0^l$	0.8070	0.8190	0.8225	0.8112	0.8029	0.8049
AUMRP$_1^l$	0.8235	0.8114	0.8087	0.8200	0.8281	0.8278

Table 2.18: AUMRP$_0^l$, AUMRP$_1^l$of GB method with $\gamma = 0.7$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $\beta = 0.1$

γ_1	**0.10**	**0.13**	**0.15**	**0.23**	**0.27**	**0.29**
n	76	80	84	100	109	115
C	57	60	64	77	85	90
P_0^*	0.6600	0.6690	0.6750	0.699	0.711	0.717
P_1^*	0.8200	0.8230	0.8250	0.833	0.837	0.839
p-value	0.0350	0.0458	0.0317	0.0458	0.0424	0.0449
power	0.9210	0.9367	0.9124	0.9358	0.9275	0.9315
AUMRP$_0^l$	0.8122	0.7966	0.8200	0.8010	0.8073	0.8047
AUMRP$_1^l$	0.8194	0.8335	0.8119	0.8319	0.8238	0.8274

Table 2.19: The AUMRP$_0^l$, AUMRP$_1^l$ of the FM method with $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\gamma_2 = 0.10$, $\alpha = 0.05$, $\beta = 0.1$

Assume that we have a sample with size n from a population who have the sensitive characteristic. The hypothesised proportion of people with the sensitive characteristic is $\pi_{A_0} = 0.7$ and the alternative proportion of people with the sensitive characteristic is $\pi_{A_1} = 0.90$.

For different values of γ_2, we determine the required minimum sample sizes and derived values of AUMRP$_0^l$. Table 2.19 investigates the AUMRP$_0^l$ and AUMRP$_1^l$ of the FM under H_0 and H_1 using Equations (2.49) and (2.50), respectively. AUMRP$_0^l$ and AUMRP$_1^l$ have values between 0.80 and 0.83. The AUMRP$_0^l$ is always greater than the AUMRP$_1^l$. However, there is no specific pattern of AUMRP$_0^l$ or AUMRP$_1^l$.

The hypothesis test based on data collected from the FM method has a p-value of approximately 0.03 for $\gamma_1 = 0.10$ which indicates that there is some evidence that the proportion of the sensitive characteristic could be not equal to 70%. Therefore, we need to use a larger sample size to get significant results. The FM method has higher power than 0.92. AUMRP_0^l of the FM method takes values between 0.80 and 0.82 whereas AUMRP_1^l of the GB method takes values between 0.81 and 0.83.

Example 2.6.3 This example derives AUMPR_0^l and AUMPR_1^l of two-sided hypothesis tests based on data collected from the GB and the FM methods. Assume that some individuals of a population have a sensitive characteristic A, with π_A as the proportion of the sensitive characteristic in a population whereas π_B is the proportion of unrelated characteristic.

This example calculates the AUMPR_0^l and AUMPR_1^l of two-sided hypothesis tests as shown in Tables 2.21 and 2.20. We calculate the required minimum sample size of the GB method using Equation (2.46) which is larger than the sample sizes of one-sided hypothesis tests, we add the continuity correction to get power larger than 0.90. The FM method needs to get a smaller sample size than the GB method to calculate AUMPR_0^l and AUMPR_1^l with higher power more than 0.90.

The proportions of the sensitive characteristic in the population of the FM method are higher than the proportions of the sensitive characteristic in the population of the GB method. This occurs due to using $\gamma > 0.7$ of the FM which is equal to $1 - \gamma_2 - \gamma_1$ whereas the GB method uses $\gamma = 0.7$. The hypothesis test based on data collected from the FM method has p-value of approximately 0.03 for $\gamma_1 = \{0.10, 0.15\}$ which indicates that there is some evidence that the proportion of the sensitive characteristic could be not equal to 70%. Therefore, we need to use a larger sample size to get significant results. Both methods have higher power than 0.90. AUMRP_0^l of the GB method takes values between 0.80 and 0.82 of the FM method takes values between 0.79 and 0.81. AUMRP_1^l of the GB method takes values between 0.80 and 0.82 of the FM method takes values between 0.81 and 0.83.

As a result, the reproducibility probability power and p-values for AUMRP_0^l and AUMRP_1^l of RRT of one-sided and two-sided tests varied depending on the threshold rejection values and the probability of yes-answer under H_0 and H_1 respectively.

π_B	**0**	**0.2**	**0.25**	**0.3**	**0.45**	**0.6**
n	149	147	147	145	140	133
l	61	69	71	72	76	78
r	84	92	94	95	98	99
P_0^*	0.490	0.550	0.565	0.580	0.625	0.670
P_1^*	0.630	0.690	0.705	0.720	0.765	0.810
p-value	0.0497	0.0466	0.0464	0.0441	0.0449	0.0430
power	0.9430	0.9427	0.9488	0.9480	0.9540	0.9620
AUMRP_0^l	0.6953	0.6993	0.7009	0.7056	0.7002	0.7013
AUMRP_1^l	0.8165	0.8142	0.8211	0.8195	0.8250	0.8334

Table 2.20: AUMRP_0^l, AUMRP_1^l using data collected from the GB method of $\gamma = 0.7$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $\beta = 0.1$.

γ_1	**0.10**	**0.13**	**0.15**	**0.23**	**0.27**	**0.29**
n	105	112	116	139	152	160
l	60	65	68	86	97	103
r	78	84	87	107	1187	125
P_0^*	0.660	0.669	0.675	0.699	0.711	0.717
P_1^*	0.820	0.823	0.825	0.833	0.837	0.839
p-value	0.0499	0.0446	0.0493	0.0426	0.0495	0.0444
power	0.9692	0.9673	0.9737	0.9664	0.9686	0.9662
AUMRP_0^l	0.6750	0.6898	0.6841	0.7023	0.6885	0.7020
AUMRP_1^l	0.8415	0.8389	0.8508	0.8394	0.8437	0.8402

Table 2.21: AUMRP_0^l and AUMRP_1^lusing data collected from the FM method of $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\gamma_2 = 0.10$ and $\alpha = 0.05$, $\beta = 0.1$

2.7 The lower and upper threshold values

This section investigates the possible range of threshold values C of each AUMRP_0^l and AUMRP_1^l. This procedure helps to find intersection points between the AUMRP_0^l and AUMRP_1^l and derivation of the lower and upper rejection threshold values $C_{\alpha,\beta}^L$ and $C_{\alpha,\beta}^U$ of threshold value C respectively. Therefore, we can obtain the same areas of AUMRP_0^l and AUMRP_1^l using different H_0 and H_1 hypotheses. In addition, these lower and upper threshold values determine the AUMRP_0^l and AUMRP_1^l with high power more than 0.90 and small p-value less than 0.05 using Equations (2.51) and (2.52) to calculate the lower

$\pi_B = 0.9$				$\pi_B = 0.8$				$\pi_B = 0.7$			
C	**AUMRP$_0^l$**	**p-value**	**power**	C	**AUMRP$_0^l$**	**p-value**	**power**	C	**AUMRP$_0^l$**	**p-value**	**power**
98	0.7811	0.0380	0.9928	95	0.7900	0.0341	0.9816	91	0.7752	0.0478	0.9791
99	0.8060	0.0223	0.9855	96	0.8135	0.0204	0.9674	92	0.7988	0.0302	0.9646
100	0.8290	0.0124	0.9725	97	0.8353	0.0116	0.9449	93	0.8212	0.0182	0.9426
101	0.8495	0.0065	0.9504	98	0.8546	0.0063	0.9110	94	0.8417	0.0105	0.9107
102	0.8668	0.0032	0.9153	99	0.8711	0.0032	0.8630	95	0.8599	0.0058	0.8666
103	0.8809	0.0015	0.8632								
$\pi_B = 0.6$				$\pi_B = 0.5$				$\pi_B = 0.4$			
C	**AUMRP$_0^l$**	**p-value**	**power**	C	**AUMRP$_0^l$**	**p-value**	**power**	C	**AUMRP$_0^l$**	**p-value**	**power**
88	0.7850	0.0414	0.9637	85	0.7949	0.0354	0.9442	81	0.7830	0.0455	0.9469
89	0.8076	0.0262	0.9427	86	0.8166	0.0224	0.9165	82	0.8047	0.0298	0.9212
90	0.8289	0.0159	0.9128	87	0.8368	0.0137	0.8792	83	0.8255	0.0188	0.8869
91	0.8482	0.0093	0.8722								
$\pi_B = 0.3$				$\pi_B = 0.2$				$\pi_B = 0.1$			
C	**AUMRP$_0^l$**	**p-value**	**power**	C	**AUMRP$_0^l$**	**p-value**	**power**	C	**AUMRP$_0^l$**	**p-value**	**power**
78	0.7937	0.0380	0.9266	74	0.7831	0.0469	0.9325	71	0.7945	0.0383	0.9123
79	0.8147	0.0247	0.8951	75	0.8044	0.0312	0.9036	72	0.8152	0.0252	0.8781
				76	0.8246	0.0201	0.8664				

Table 2.22: AUMPR$_0^l$ of the GB method versus C with $n = 119$, $\pi_{A_0} = 0.7$, $\gamma = 0.7$, $\pi_{A_1} = 0.9$, with the corresponding p-value and power

and upper rejection threshold values $C^L_{\alpha,\beta}$ and $C^U_{\alpha,\beta}$ for each value $\pi_B \in [0, 1]$, using the minimum required sample size n as follows:

$$P(Y^{2n}_{n+1} \leq C^L_{\alpha,\beta} \mid P^* = P^*_0) = \sum_{y=0}^{C^L_{\alpha,\beta}} \binom{n}{y} (P^*_0)^y (1 - P^*_0)^{n-y} \leq 1 - \alpha \tag{2.51}$$

$$P(Y^{2n}_{n+1} > C^U_{\alpha,\beta} \mid P^* = P^*_1) = 1 - \sum_{y=0}^{C^U_{\alpha,\beta}} \binom{n}{y} (P^*_1)^y (1 - P^*_1)^{n-y} \geq 1 - \beta \tag{2.52}$$

Equation (2.51) derive the range of values of the event $Y = y$ in which the probability of type of error I under H_0 while Equation (2.52) derive the range of values of the event $Y = y$ in which the probability of type of error II under H_1.

All the integer values between rejection threshold values $C^L_{\alpha,\beta}$ and $C^U_{\alpha,\beta}$ are determined to derive AUMRP$_0^l$ and AUMRP$_1^l$. Then, the p-value and the power for each rejection threshold $C \in [C^L_{\alpha,\beta}, C^U_{\alpha,\beta}]$. Therefore, H_0 is not rejected if $Y^{2n}_{n+1} \leq C^L_{\alpha,\beta}$ and is rejected if

$\pi_B = 0.9$				$\pi_B = 0.8$				$\pi_B = 0.7$			
C	**AUMRP$_1^l$**	**p-value**	**power**	C	**AUMRP$_1^l$**	**p-value**	**power**	C	**AUMRP$_1^l$**	**p-value**	**power**
98	0.8783	0.0380	0.9928	95	0.8556	0.0341	0.9816	91	0.8533	0.0478	0.9791
99	0.8603	0.0223	0.9855	96	0.8344	0.0204	0.9674	92	0.8327	0.0302	0.9646
100	0.8387	0.0124	0.9725	97	0.8101	0.0116	0.9449	93	0.8096	0.0182	0.9426
101	0.8133	0.0065	0.9504	98	0.7833	0.0063	0.9110	94	0.7842	0.0105	0.9107
102	0.7846	0.0032	0.9153	99	0.7549	0.0032	0.8630	95	0.7577	0.0058	0.8666
103	0.7538	0.0015	0.8632								
$\pi_B = 0.6$				$\pi_B = 0.5$				$\pi_B = 0.4$			
C	**AUMRP$_1^l$**	**p-value**	**power**	C	**AUMRP$_1^l$**	**p-value**	**power**	C	**AUMRP$_1^l$**	**p-value**	**power**
88	0.8329	0.0414	0.9637	85	0.8131	0.0354	0.9442	81	0.8163	0.0455	0.9469
89	0.8107	0.0262	0.9427	86	0.7901	0.0224	0.9165	82	0.7942	0.0298	0.9212
90	0.7867	0.0159	0.9128	87	0.7660	0.0137	0.8792	83	0.7710	0.0188	0.8869
91	0.7615	0.0093	0.8722								
$\pi_B = 0.3$				$\pi_B = 0.2$				$\pi_B = 0.1$			
C	**AUMRP$_1^l$**	**p-value**	**power**	C	**AUMRP$_1^l$**	**p-value**	**power**	C	**AUMRP$_1^l$**	**p-value**	**power**
78	0.7990	0.0380	0.9266	74	0.8043	0.0469	0.9325	71	0.7887	0.0383	0.9123
79	0.7766	0.0247	0.8951	75	0.7825	0.0312	0.9036	72	0.7669	0.0252	0.8781
				76	0.7601	0.0201	0.8664				

Table 2.23: AUMPR$_1^l$ of the GB method versus C with $n = 119$, $\pi_{A_0} = 0.7$, $\gamma = 0.7$, $\pi_{A_1} = 0.9$, with the corresponding p-value and power

$Y_{n+1}^{2n} > C_{\alpha,\beta}^U$. Using the assumptions of Examples 2.6.2 and 2.6.1, we explain this range and the effect of changing the parameters in detail.

Figures 2.20 and 2.21 show the AUMRP$_0^l$ and AUMRP$_1^l$ of GB and FM methods in comparison to an extended range of rejection threshold values C using the largest minimum sample size to determine the rejection threshold $C \in [C_{\alpha,\beta}^L, C_{\alpha,\beta}^U]$. It is clear that there are intersection points between AUMRP$_0^l$ and AUMRP$_1^l$ for $P^* = P_0^*$. It is obvious that AUMRP$_0^l$ and AUMRP$_1^l$ are equal if they have the same threshold or probability P_0^*, which occurs at the intersection of them.

As shown in Tables 2.23 and 2.22, AUMPR$_0^l$ and AUMPR$_1^l$ of the GB method. The highest threshold values (98, 103) are determined for the GB methods at $\pi_B = 0.9$ and then AUMPR$_0^l$ takes values between 0.78 and 0.88. For the same threshold value, AUMPR$_1^l$ takes values between 0.75 and 0.87 with a power decrease from 0.0380 to 0.0015. The lowest threshold (71, 72) values are determined for the GB methods at $\pi_B = 0.1$ and then

$\gamma_1 = 0.29$				$\gamma_1 = 0.26$				$\gamma_1 = 0.23$			
C	**AUMRP**$_1^0$	**p-value**	**power**	C	**AUMRP**$_1^0$	**p-value**	**power**	C	**AUMRP**$_1^0$	**p-value**	**power**
90	0.8047	0.0449	0.9315	89	0.8130	0.0458	0.9485	88	0.8029	0.0467	0.9617
91	0.8336	0.0276	0.8945	90	0.8411	0.0284	0.9187	89	0.8311	0.0292	0.9382
				91	0.8678	0.0168	0.8771	90	0.8582	0.0174	0.9043
								91	0.8833	0.0099	0.8580
$\gamma_1 = 0.15$				$\gamma_1 = 0.13$				$\gamma_1 = 0.10$			
C	**AUMRP**$_1^0$	**p-value**	**power**	C	**AUMRP**$_1^0$	**p-value**	**power**	C	**AUMRP**$_1^0$	**p-value**	**power**
86	0.8195	0.0362	0.9764	85	0.8100	0.0425	0.9842	84	0.8095	0.0430	0.9887
87	0.8463	0.0224	0.9608	86	0.8369	0.0268	0.9729	85	0.8362	0.0273	0.9803
88	0.8716	0.0133	0.9373	87	0.8627	0.0162	0.9555	86	0.8617	0.0166	0.9669
89	0.8947	0.0075	0.9037	88	0.8865	0.0093	0.9297	87	0.8854	0.0097	0.9466
90	0.9153	0.0041	0.8579	89	0.9079	0.0052	0.8934	88	0.9067	0.0054	0.9172
								89	0.9254	0.0029	0.8765
$\gamma_1 = 0.08$				$\gamma_1 = 0.04$				$\gamma_1 = 0.00$			
C	**AUMRP**$_1^0$	**p-value**	**power**	C	**AUMRP**$_1^0$	**p-value**	**power**	C	**AUMRP**$_1^0$	**p-value**	**power**
84	0.8270	0.0322	0.9869	82	0.8087	0.0439	0.9945	81	0.8171	0.0385	0.9958
85	0.8528	0.0199	0.9773	83	0.8350	0.0281	0.9899	82	0.8428	0.0244	0.9923
86	0.8770	0.0118	0.9624	84	0.8601	0.0173	0.9824	83	0.8672	0.0149	0.9864
87	0.8991	0.0067	0.9400	85	0.8834	0.0102	0.9703	84	0.8896	0.0088	0.9767
88	0.9186	0.0037	0.9080	86	0.9045	0.0058	0.9519	85	0.9098	0.0049	0.9617
89	0.9356	0.0019	0.8644	87	0.9231	0.0031	0.9251	86	0.9275	0.0027	0.9394
				88	0.9392	0.0016	0.8877	87	0.9427	0.0014	0.9077
								88	0.9555	0.0007	0.8646

Table 2.24: AUMPR$_0^l$ of the FM method versus C with $n = 115$ with the corresponding p-value and power

AUMPR$_0^l$ takes values between 0.79 and 0.81. For the same threshold value, AUMPR$_1^l$ takes values between 0.76 and 0.77 with a power decrease from 0.0380 to 0.0015.

Tables 2.24 and 2.25 show AUMPR$_0^l$ and AUMPR$_1^l$ of the FM method. The highest threshold values between 91, and 90 that are determined for the GB methods at $\gamma_1 = 0.29$ and then AUMPR$_0^l$ takes values between 0.79 and 0.82. For the same threshold value, AUMPR$_1^l$ takes values between 0.78 and 0.82 with power decrease from 0.9315 to 0.8945 and p-value decrease from 0.0449 to 0.0276. The lowest threshold $(81,\ 88)$ values are determined for the GB methods at $\gamma_1 = 0.0$ and then AUMPR$_0^l$ takes values between 0.95 and 0.81. For the same threshold value, AUMPR$_1^l$ takes values between 0.93 and 0.77 with power decrease from 0.99 to 0.86 and p-value decrease from 0.0385 to 0.0007.

$\gamma_1 = 0.29$				$\gamma_1 = 0.26$				$\gamma_1 = 0.23$			
C	**AUMRP**$^l_1$	**p-value**	**power**	C	**AUMRP**$^l_1$	**p-value**	**power**	C	**AUMRP**$^l_1$	**p-value**	**power**
90	0.8274	0.0449	0.9315	89	0.8411	0.0458	0.9485	88	0.8611	0.0467	0.9617
91	0.7974	0.0276	0.8945	90	0.8122	0.0284	0.9187	89	0.8338	0.0292	0.9382
				91	0.7819	0.0168	0.8771	90	0.8045	0.0174	0.9043
								91	0.7742	0.0099	0.8580
$\gamma_1 = 0.15$				$\gamma_1 = 0.13$				$\gamma_1 = 0.10$			
C	**AUMRP**$^l_1$	**p-value**	**power**	C	**AUMRP**$^l_1$	**p-value**	**power**	C	**AUMRP**$^l_1$	**p-value**	**power**
86	0.8843	0.0362	0.9764	85	0.9003	0.0425	0.9842	84	0.9121	0.0430	0.9887
87	0.8599	0.0224	0.9608	86	0.8782	0.0268	0.9729	85	0.8919	0.0273	0.9803
88	0.8330	0.0133	0.9373	87	0.8532	0.0162	0.9555	86	0.8687	0.0166	0.9669
89	0.8041	0.0075	0.9037	88	0.8257	0.0093	0.9297	87	0.8428	0.0097	0.9466
90	0.7742	0.0041	0.8579	89	0.7966	0.0052	0.8934	88	0.8148	0.0054	0.9172
								89	0.7854	0.0029	0.8765
$\gamma_1 = 0.08$				$\gamma_1 = 0.04$				$\gamma_1 = 0.00$			
C	**AUMRP**$^l_1$	**p-value**	**power**	C	**AUMRP**$^l_1$	**p-value**	**power**	C	**AUMRP**$^l_1$	**p-value**	**power**
84	0.9071	0.0322	0.9869	82	0.9325	0.0439	0.9945	81	0.9393	0.0385	0.9958
85	0.8861	0.0199	0.9773	83	0.9159	0.0281	0.9899	82	0.9240	0.0244	0.9923
86	0.8622	0.0118	0.9624	84	0.8964	0.0173	0.9824	83	0.9060	0.0149	0.9864
87	0.8358	0.0067	0.9400	85	0.8740	0.0102	0.9703	84	0.8850	0.0088	0.9767
88	0.8074	0.0037	0.9080	86	0.8489	0.0058	0.9519	85	0.8614	0.0049	0.9617
89	0.7780	0.0019	0.8644	87	0.8216	0.0031	0.9251	86	0.8352	0.0027	0.9394
				88	0.7927	0.0016	0.8877	87	0.8071	0.0014	0.9077
								88	0.7781	0.0007	0.8646

Table 2.25: AUMPR$^l_1$ of the FM method versus C with $n = 115$ with the corresponding p-value and power

In general, the power increases and the p-value decreases if π_B and γ_1 decrease. That happens because of the range of the threshold values decreases if π_B increases whereas the range of the threshold values increases if γ_1 decreases.

In addition, under H_0, increasing the value C from the lower threshold value to the upper threshold value for each π_B or γ_1 values leads to decreasing the power and p-values and then leads to increasing AUMRP$^l_0$ of RRT. Conversely, under H_1, increasing the value C from the lower threshold value to the upper threshold value leads to decreasing the power and p-values and decreasing AUMRP$^l_1$ of RRT.

Furthermore, increasing the value C under H_0 causes the power and p-values to decrease

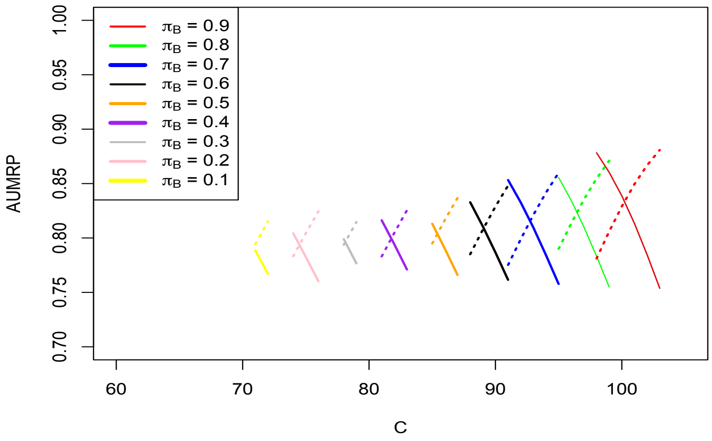


Figure 2.20: The AUMRP$_0^l$ with solid lines and AUMRP$_1^l$ with dotted lines of GB method versus C with $n = 119$, $\pi_{A_0} = 0.7$, $\gamma = 0.7$, $\pi_{A_1} = 0.9$

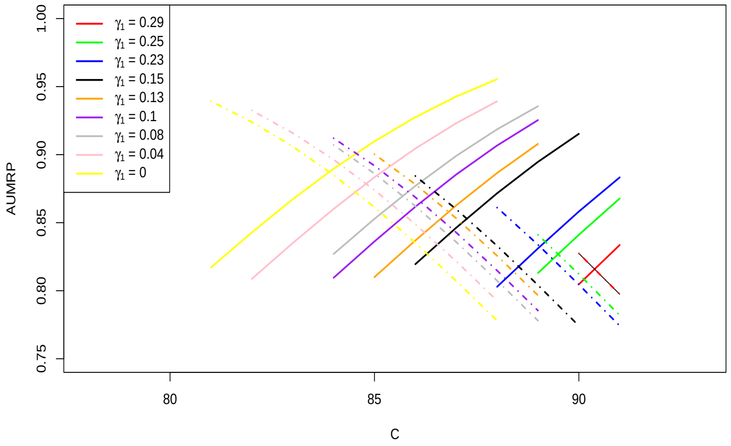


Figure 2.21: The AUMRP$_0^l$ with solid lines and AUMRP$_1^l$ with dotted lines of the FM method versus C with $n = 115$, $\pi_{A_0} = 0.7$, $\pi_{A_1} = 0.9$, $\gamma_1 = 0.10$

and the AUMRP$_0^l$ of RRT to increase as it moves from the lower threshold value to the upper threshold value. Conversely, when C is increased from the lower threshold value to the upper threshold value under H_1, the power and p-values decrease along with AUMRP$_1^l$ of RRT.

2.8 Comparison of the reproducibility of statistical tests based on different RRT methods

In this section, we compare RRT methods considering the variance of the estimators and reproducibility of statistical hypothesis tests using the same privacy degree.

We assume a large sample size for the GB and FM methods and choose different parameters for the RRT methods to get the same privacy and the variance of the estimator $\hat{\pi}_{A_0}$ because of larger sample sizes usually increase the reproducibility probability occurring inside the $[0, 1]$.

This choice of the parameters gives the same values of both variances of the estimator $\hat{\pi}_{A_0}$ and the same privacy degree of the GB and FM method to check the changes in reproducibility of statistical hypothesis tests.

Example 2.8.1 Assume that we have $n = 500$, $\gamma = 0.5554$, $\pi_{A_1} = 0.9$ as parameters of the GB method and $\gamma_1 = 0.20829$, $\gamma_2 = 0.10$, $\pi_{A_1} = 0.9$, $\alpha = 0.05$, $\beta = 0.1$ as parameters of the FM method.

We compare the GB and FM methods for all values of $\pi_{A_0} = \{0.550, 0.555, 0.560, 0.565, 0.570, 0.575, 0.580, 0.585, 0.590, 0.595, 0.600\}$ in terms of the reproducibility when both methods have the same privacy degree around 1.233 and the same variance of the estimator $\hat{\pi}_{A_0}$ as shown in Tables 2.26 and 2.27.

For the GB method with privacy degree $\Delta_{GB} = 1.2237$, Table 2.26 shows lower variance for different values of $\hat{\pi}_{A_0}$because the changes in $\hat{\pi}_{A_0}$ values are very small about 0.005. The AUMRP$_0^l$ and AUMRP$_1^l$ have no pattern while the power is always high around 1.0000. AUMRP$_0^l$ takes values between 0.82 and 0.83 whereas AUMRP$_0^l$ takes values around 1.0000 for different π_{A_0}.

For the FM method with privacy degree, $\Delta_{FM} = 1.2236$, Table 2.27 shows lower variance for different values of π_{A_0} and there is no big difference between these variances. The AUMRP$_0^l$ and AUMRP$_1^l$ have no pattern while the power is always high about 1.0000.

π_{A_0}	0.550	0.555	0.560	0.565	0.570	0.575	0.580	0.585	0.590	0.595	0.600
$\text{Var}(\hat{\pi}_{A_0})_{GB}$	0.00162	0.00162	0.00162	0.00162	0.00162	0.00162	0.00162	0.00162	0.00162	0.00162	0.00162
Power	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AUMRP_0^l	0.8274	0.8230	0.8298	0.8255	0.8323	0.8280	0.8237	0.8306	0.8263	0.8331	0.8288
AUMRP_1^l	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2.26: The $\text{Var}(\hat{\pi}_{A_0})_{GB}$, Δ_{GB} , AUMRP_0^l, AUMRP_1^l of one-sided tests based on GB data of $n = 500$, $\pi_B = 0.4$, $\pi_1 = 0.9$, $\gamma = 0.5554$, $\alpha = 0.05$, $\beta = 0.1$, $\Delta_{GB} = 1.2237$

π_{A_0}	0.550	0.555	0.560	0.565	0.570	0.575	0.580	0.585	0.590	0.595	0.600
$\text{Var}(\hat{\pi}_{A_0})_{FM}$	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002	0.00002
Power	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
AUMRP_0^l	0.8322	0.8240	0.8335	0.8301	0.8331	0.8250	0.8281	0.8312	0.8230	0.8261	0.8312
AUMRP_l^l	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 2.27: The $\text{Var}(\hat{\pi}_{A_0})_{FM}$, Δ_{FM} , AUMRP_0^l, AUMRP_1^l of one-sided tests based on FM data of $n = 500$, $\gamma_2 = 0.10$, $\pi_{A_1} = 0.9$, $\gamma_1 = 0.20829$, $\alpha = 0.05$, $\beta = 0.1$, $\Delta_{FM} = 1.2236$

AUMRP_0^l takes values between 0.82 and 0.83 whereas AUMRP_0^l takes values 1.0000 for different π_{A_0}.

In general, a larger sample size leads to higher reproducibility and larger areas of AUMRP_0^l, AUMRP_1^l. For example, $\text{AUMRP}_0^l = 0.824$ for $n = 500$ and $\pi_{A_0} = 0.555$ whereas $\text{AUMRP}_0^l = 0.7873$ for $n = 30$ and $\pi_{A_0} = 0.555$.

The FM method has a lower variance than the GB methods whereas AUMRP_0^l of the FM method get higher reproducibility than the GB method π_{A_0} for the same privacy degree 1.2236.

As shown in Tables 2.26 and 2.27, even though the estimator's variances are extremely low in order to achieve higher reproducibility, however, it might not be practical to assume the mentioned parameters in order to reduce privacy degree. Therefore, it is better to suppose different hypothesised values and other parameters to obtain the same degree of privacy with less variability in the true responses and higher reproducibility.

2.9 Concluding remarks

This chapter has presented a novel method for determining the reproducibility probability of statistical hypothesis tests based on data collected by RRT methods, such as the GB and the FM methods. This method uses the number of 'Yes' responses for a particular data set and the testing threshold. Then we apply NPI for Bernoulli quantities to compute the lower and upper probabilities of an event in one-sided and two-sided tests.

For reproducibility of one-sided hypothesis tests, we introduce the measurement of lower and upper reproducibility probability MRP_0^l and MRP_0^u under H_0 using a single threshold value. Similarly, we introduce the measurement of lower and upper reproducibility probability MRP_1^l and MRP_1^u under H_1 respectively. Then we compare the GB and the FM methods by derivation of the required minimum sample size with respect to higher power more than 0.90 and p-value less than 0.05. After that, we calculate the area under MRP_0^l and MRP_1^l. In addition, we derive the lower and upper threshold values to find the same area of the threshold value of MRP_0^l and MRP_1^l using different parameters of RRT method and using the largest minimum sample sizes and with respect to higher power more than 0.90 and p-value less than 0.05. The FM method has more reproducibility than the GB methods for small sizes for the same parameters for one-sided tests.

For larger sample sizes $n = 500$, the same variance, privacy degree, and proportion of sensitive characteristics in the population π_{A_0}, the FM method has higher reproducibility than the GB method. The FM method takes smaller samples than the GB method requires. As a result, choosing the same parameters within p-value less than 0.05 and power larger than 0.90 need to increase the sample size of the GB method than the FM method to obtain the same AUMRP_0^l and AUMRP_1^l for two-sided tests.

The FM method has higher reproducibility than the GB method for larger sample sizes $n = 500$ for the same variance, the same degree of privacy, and the same proportion of sensitive characteristics in the population π_{A_0} using one-sided hypothesis tests. This occurs because the FM method's $(1-\gamma_1-\gamma_2)$ is larger than the GB method's $(1-\gamma)$. Furthermore, the FM approach requires smaller samples than the GB method. Therefore, using the same sample size for both methods results in improved reproducibility of hypothesis tests of the FM method.

So, we find that less variability in the reported responses of any RRT method leads to higher reproducibility with the same degree of privacy.

The advantage of employing reproducibility of statistical tests is that they can be designed for any RRT method. It can be applied to Warner data to compare with the Greenberg method, however, this comparison has been done if we assume the Warner method is a special case from the Greenberg method when $1 - \pi_A$ in the WM is equal to π_B in the GB.

For the limitations of this method, there is practically no reason why our reproducibility method (NPI-RP) cannot be used with larger sample quantities. The principle of the NPI-B method can be applied to any sample using the sample orderings method. The number of orderings need to be sampling not depend on the sample size but depends on the binomial distribution. Therefore, we can apply this method depending on any sample size. If someone ran into computational problems with larger sample sizes, it was probably just a software or computer issue.

Chapter 3

Reproducibility of estimates

3.1 Introduction

Estimation of population characteristics is an essential part of statistical inference. In this chapter, we investigate the reproducibility of estimates. However, it is clear that for real-valued random quantities, an estimate of a parameter or the characteristic will not be reproduced precisely. Therefore, we define reproducibility of estimate as the probability of the event that, if we repeat the experiment under the same circumstances, the estimate based on the future sample will be close to the estimate based on the original sample.

The objective of this chapter is to introduce NPI for reproducibility of estimates using two procedures. The first procedure is reproducibility of estimates using NPI-B method. The second procedure is reproducibility of estimates using a representative sample of a population, which is a novel concept we introduce here without making any further assumptions. We investigate the reproducibility of estimates of population characteristics such as the mean, median, variance, quartiles and interquartile ranges.

This chapter is structured as follows. Section 3.2 introduces reproducibility of estimates, in general, using NPI-B method. In Section 3.3, the concept of a representative sample of the underlying population distribution is introduced, together with its use to asses reproducibility of estimates and a comparison of both techniques. Section 3.4 presents some concluding remarks.

3.2 Reproducibility of estimates using NPI-B method

In this section, a new theory of reproducibility for estimates is proposed. Suppose that we have n real-valued random quantities $X_1, X_2, ..., X_n$, which are assumed to be independent and identically distributed. Let's assume that the ordered observed values of these random quantities be denoted by $x_1 < ... < x_n$. For simplicity of implementing this theory, we determine the lower and upper bounds of these random quantities which are x_0 and x_{n+1} to avoid possible probability mass of $-\infty$ or ∞ that could impact the mean of the future m observations. These bounds can be specified using

$$x_0 = \min_{1 \leq i \leq n}(x_i) - d, \quad x_{n+1} = \max_{1 \leq i \leq n}(x_i) + d \tag{3.1}$$

where d is the maximum distance between two consecutive observations.

The n observations can be divided the real line into $n+1$ intervals, which are $I_i = (x_{i-1}, x_i)$ for $i = 1, ..., n+1$. Assume that the estimate based on the original sample is $\hat{\theta}$, and the estimate based on the future sample is $\hat{\theta}^f$ where $\hat{\theta}^f = \hat{\theta} \pm \epsilon$ and ϵ is the distance between the two estimates and takes values $\epsilon \geq 0$. As a result, we call this theory $\epsilon-$reproducibility for estimates

In this method, we assume that there are no ties between the original and the future observations for simplicity.

Assume that an original estimate $\hat{\theta}$ of the original sample and an estimate based on a future data set $\hat{\theta}^f$ should be in $[\hat{\theta} - \epsilon, \ \hat{\theta} + \epsilon]$. The probability of the event $|\hat{\theta} - \hat{\theta}^f| \leq \epsilon$, which is defined as the probability of the absolute value of the difference between the original estimate and the future estimate which is equal or less than any real value ϵ, is used to derive the reproducibility for an estimate as follows:

$$RP(\epsilon) = P(|\hat{\theta} - \hat{\theta}^f| \leq \epsilon) \tag{3.2}$$

where $\hat{\theta}$ estimate of a population characteristic.

To illustrate this method to quantify of reproducibility of estimates using the method of NPI-B method for observations on finite intervals as follows. Based on the original sample $x_1, ..., x_n$, we estimate the population characteristic θ by $\hat{\theta}$ to assess $\epsilon-$reproducibility of this estimate. We use NPI-B method as explained in Section 1.6 to create future samples of size m where $m = n$. For such future samples, we also derive the estimate of θ denoted by $\hat{\theta}^f$ and this allows us to estimate $\epsilon-$reproducibility. For simplicity of implementing the NPI-B method, we assume finite support of X_i using Equation (3.1).

NPI-B method, as described in Section 1.6, is used to generate a future sample $b_1, ..., b_n$ and denote the estimate based on this bootstrap sample by $\hat{\theta}_{B_i}$. We perform this procedure n_B times. NPI-B method draws new observations from the whole range of possible observations and outside the bounds of this original sample.

Based on these n_B bootstrap samples, we can estimate reproducibility of $\hat{\theta}$ by:

$$\hat{RP}(\epsilon) = \sum_{i=1}^{n_B} \frac{1}{n_B} \quad \mathbf{1}\Big\{ |\hat{\theta} - \hat{\theta}_{B_i}| \leq \epsilon \Big\} \tag{3.3}$$

with $\epsilon \geq 0$, and $\mathbf{1}\{\mathbf{A}\}$ is an indicator function that is equal to 1 if event $\mathbf{A}$ is true and 0 otherwise.

Note that, by using NPI-B method, we get a precise of $\hat{RP}(\epsilon)$, so there is no imprecision and repeated applications of this bootstrap will leads to different estimate $\hat{RP}(\epsilon)$. In the following example, we illustrate this procedure.

Example 3.2.1 To illustrate reproducibility for an estimate using NPI-B method for two different samples. We assume that we have the first sample with size $n = 30$ from the standard normal distribution:

$\mathrm{X}_i = \{-1.8180, -1.5977, -1.5531, -0.9193, -0.8864, -0.7505, -0.6443, -0.4816, -0.4535,$
$-0.3316, -0.2842, -0.2762, -0.1623, -0.1162, -0.1093, 0.2987, 0.3706, 0.5202, 0.5855, 0.6059,$
$0.6121, 0.6204, 0.6301, 0.7095, 0.7796, 0.8169, 1.1207, 1.4558, 1.8051, 1.8173\}$

The sample mean $\bar{x} = 0.0788$ is an estimate of the population mean μ and we are interested in $\epsilon-$reproducibility of this estimate based on these data and the suggestion in

n_B	$\max(\bar{x}_B)$	$\text{mean}(\bar{x}_B)$	$\min(\bar{x}_B)$	$RP(\epsilon)$
100	0.5946	0.0863	-0.6157	1.0000
500	0.8046	0.0799	-0.8492	1.0000
1000	0.8353	0.0795	-0.8492	1.0000
10000	0.9730	0.0765	-1.0055	1.0000
100000	1.1146	0.0757	-1.1764	1.0000

Table 3.1: $\hat{RP}(\epsilon)$ of the mean of characteristics of the standard normal distribution of the first sample with $n = 30$, $\epsilon = 1$

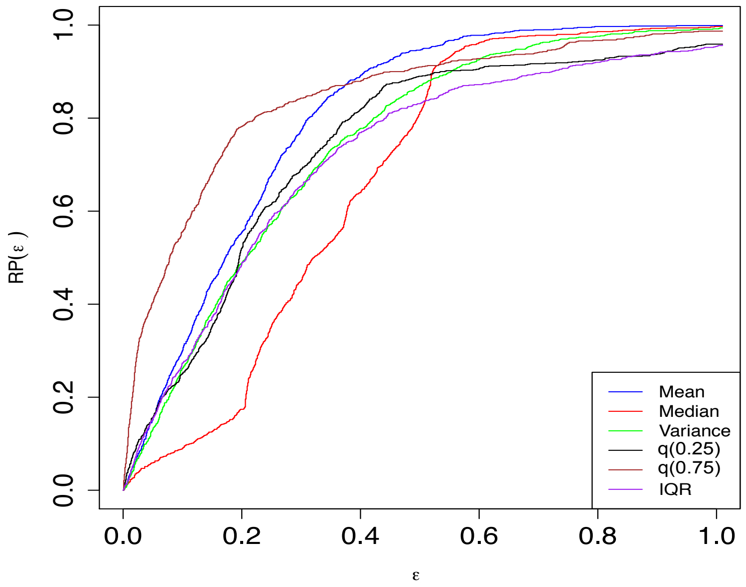


Figure 3.1: $\hat{RP}(\epsilon)$ for estimates of characteristics of the standard normal distribution of the first sample with $n = 30,\ n_B = 1000$

Equations (3.1) with the maximum distance between consecutive observations $d = 0.6338$. We set the lower and upper bounds of the support, used in the NPI-B method, of $x_0 = -2.4518$ and $x_{n+1} = 2.4511$.

Figure 3.1 shows the $\epsilon-$reproducibility for estimates of characteristics of the standard normal distribution with sample of size $n = 30$ and bootstrap numbers $n_B = 1000$. These characteristics are the mean with the blue line, the median with the red line, variance with the green line, the first quartile with the black line, the third quartile with the dark red line and the IQR with the purple line.

It is noted that $\hat{RP}(\epsilon)$ increases for all characteristics if ϵ increases. The $\epsilon-$reproducibility

n_B	$\max(\bar{x}_B)$	$\text{mean}(\bar{x}_B)$	$\min(\bar{x}_B)$	$\hat{RP}(\epsilon)$
100	0.2348	0.0851	-0.0613	1.0000
500	0.2758	0.0859	-0.0955	1.0000
1000	0.2758	0.0818	-0.1057	1.0000
10000	0.3142	0.0827	-0.1507	1.0000
100000	0.3814	0.0824	-0.1585	1.0000

Table 3.2: $\hat{RP}(\epsilon)$ of the mean of characteristics of the standard normal distribution of the first sample with $n = 500$, $\epsilon = 1$

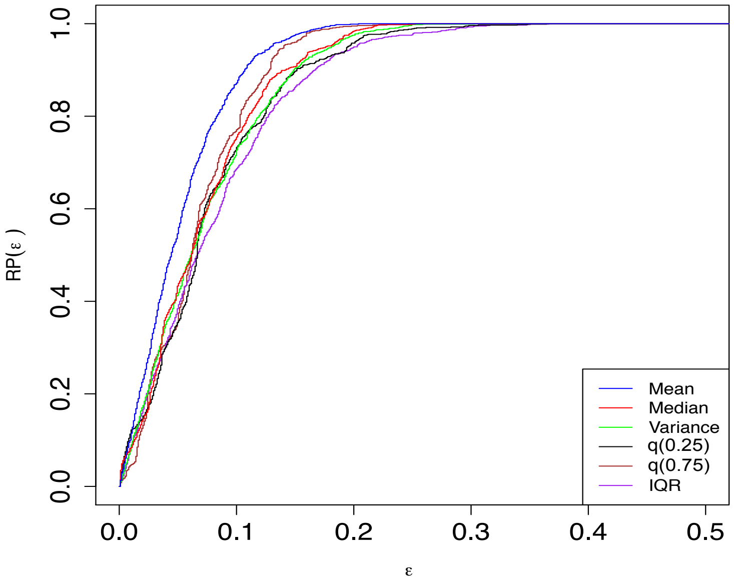


Figure 3.2: $\hat{RP}(\epsilon)$ for estimates of characteristics of the standard normal distribution of the first sample with $n = 500$, $n_B = 1000$

of the median takes the lowest values for $\epsilon \in [0, 0.5]$ whereas the reproducibility of the third quartiles take the largest values for $\epsilon \in [0, 0.3]$. The $\epsilon-$reproducibility of other characteristic lines show fluctuation for $\epsilon \in [0, 0.5]$. For the values $\epsilon \in (0.5, 1]$, the $\epsilon-$reproducibility of the mean takes the highest values whereas the lowest values of $\epsilon-$reproducibility for estimates fluctuate between values of $\epsilon-$reproducibility of $q(0.25)$ and IQR.

We generate $n_B = 1000$ NPI-B samples leading 1000 estimates $\bar{x}_{B_i}$for the population mean μ, where $i = 1, ..., 1000$. The resulting estimates of $\hat{RP}(\epsilon)$, for $\epsilon \in [0, 1]$, are presented by the various colored lines in Figure 3.1.

For more variations, we repeat the bootstrap with different times of n_B using the same

n_B	$\max(\bar{x}_B)$	$\text{mean}(\bar{x}_B)$	$\min(\bar{x}_B)$	$\hat{RP}(\epsilon)$
100	1.0002	0.0504	-0.4222	1.0000
500	1.6648	0.0412	-0.9998	0.9900
1000	1.6648	0.0637	-1.0122	0.9870
10000	1.7508	0.0747	-1.0515	0.9905
100000	2.1930	0.0739	-1.3787	0.9908

Table 3.3: $\hat{RP}(\epsilon)$ of the mean of characteristics of the standard normal distribution of the second sample with $n = 30$, $\epsilon = 1$

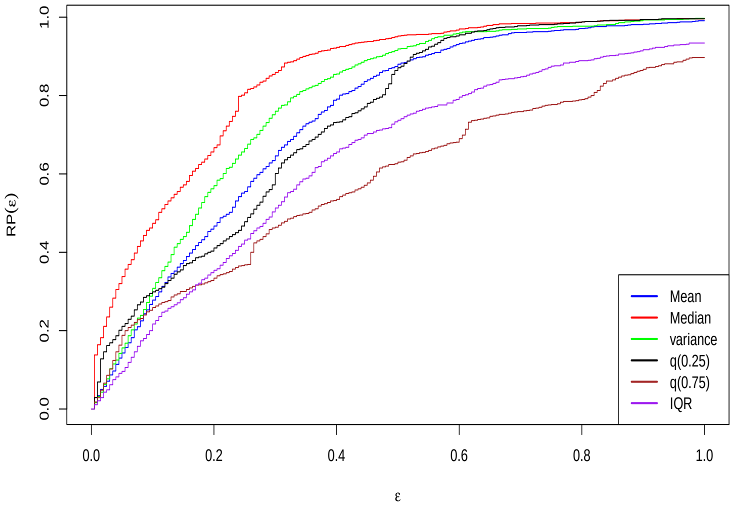


Figure 3.3: $\hat{RP}(\epsilon)$ for estimates of characteristics of the standard normal distribution of the second sample with $n = 30,\ n_B = 1000$

sample to get the $\epsilon-$reproducibility for the mean as shown in Table 3.1. It is noted that the mean of the original sample is 0.0788 and the closest value of mean of the bootstrap sample means is 0.0765 of $n_B = 10000$. In addition, increasing the n_B to 100000 leads to a fixed value of the $\epsilon-$reproducibility of the mean and the mean of the bootstrap sample means.

Based on the same sample and the same bootstrap samples, we have also estimated the $\epsilon-$reproducibility of estimates of the population median, variance, the first and the third quartiles (denoted by $q(0.25)$ and $q(0.75)$, respectively) and the inter-quartile range. These estimates are also presented in Figure 3.1.

n_B	max($\bar{x}_B$)	mean($\bar{x}_B$)	min($\bar{x}_B$)	$\hat{RP}(\epsilon)$
100	0.1341	-0.0466	-0.2167	1.0000
500	0.1948	-0.0482	-0.2669	1.0000
1000	0.1948	-0.0526	-0.2959	1.0000
10000	0.2139	-0.0520	-0.3042	1.0000
100000	0.2647	-0.0523	-0.3253	1.0000

Table 3.4: $\hat{RP}(\epsilon)$ of the mean of characteristics of the standard normal distribution of a different sample with $n = 500$, $\epsilon = 1$

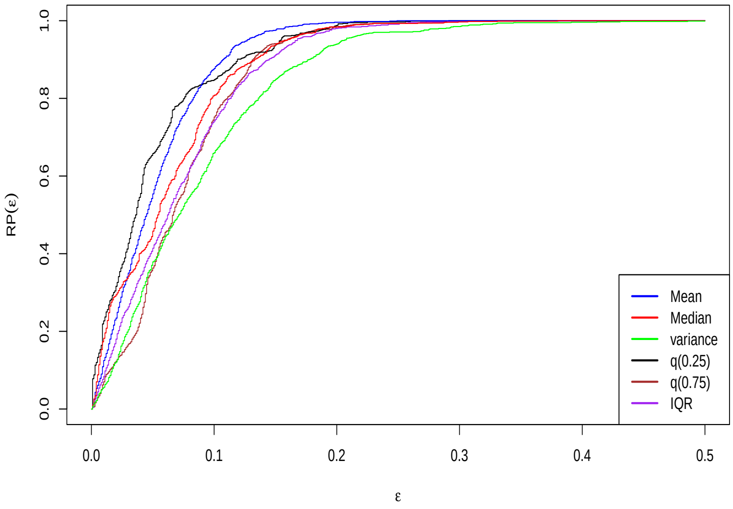


Figure 3.4: $\hat{RP}(\epsilon)$ for estimates of characteristics of the standard normal distribution of a different sample with $n = 500$, $n_B = 1000$

As shown in Figure 3.2, the $\epsilon-$reproducibility of the mean for a similar example but with sample size $n = 500$ with mean $\bar{x} = 0.0825$ where $x_0 = -2.7834$ and $x_{n+1} = 2.9489$. This illustrates that $\epsilon-$reproducibility is much better for a larger sample sizes as expected because of reduction of the variability of the estimates.

Now, we illustrate the $\epsilon-$reproducibility for the mean using with the same procedure and using different samples to derive the $\epsilon-$reproducibility of the mean for a sample of size $n = 30$ from the standard normal distribution:

$\text{X}_i = \{0.3408, -0.7033, -0.3795, -0.7460, -0.8981, -0.3348, -0.5014, -0.1745, 1.8090, -0.2301, -1.1304, 0.2160, 1.2322, 1.6094, 0.4016, -0.2730, -0.0362, -0.1503, 3.7688, -1.6525,$

$-1.1351, 0.2277, -0.1833, -0.4135, -0.4376, -0.0262, -0.8598, 0.1665, 1.4755, 0.1954\}$

with mean $\bar{x} = 0.0392$ where $x_0 = -3.6123$. $x_{n+1} = 5.7286$ and $d = 1.9598$. As can be shown in Figure 3.3, the $\epsilon-$reproducibility of median takes the highest values whereas the $\epsilon-$reproducibility of IQR takes the lowest values for $\epsilon \in [0, 0.18]$ and the $\epsilon-$reproducibility of $q(0.75)$ takes the lowest values for $\epsilon \in (0.18, 1]$. The difference between the $\epsilon-$reproducibility of characteristics of this sample is clear because of the second sample has fluctuating the $\epsilon-$reproducibility of estimates more than the first sample. In addition, the $\epsilon-$reproducibility of estimates becomes much better if we increase the sample to $n = 500$ as shown in Figure 3.4.

For more variations, we use different bootstrap samples as shown in Tables 3.3 and 3.4. It is shown that the $\epsilon-$reproducibility of estimates of the first sample is higher than the $\epsilon-$reproducibility of estimates of the second sample. Similarly, the mean of the original samples and the means of bootstrap sample means of the first sample is higher than the mean of the original samples and the means of bootstrap sample means of the second sample although both samples are generated from the standard normal distribution.

Example 3.2.2 In this example, we do this with same procedure and using different distribution to derive the $\epsilon-$reproducibility of the mean. Assume that we have a sample of size $n = 30$ from the exponential distribution with rate $\lambda = 5$, where

$\mathrm{X}_i = \{0.4891, 0.0184, 0.5006, 0.0063, 0.2524, 0.2318, 0.6305, 0.2023, 0.5623, 0.0245,$
$0.2767, 0.0873, 0.1869, 0.4327, 0.2067, 0.1149, 0.1792, 0.0339, 0.1182, 0.3984, 0.0008, 0.1527,$
$0.4234, 0.0426, 0.4057, 0.0204, 0.0154, 0.3222, 0.2767, 0.2563\}$.

The sample mean $\bar{x} = 0.2364$ is an estimate of the population mean μ and we are interested in the $\epsilon-$reproducibility of this estimate based on these data and the suggestion in Equations (3.1) with the maximum distance between consecutive observations $d = 0.5164$. We set the lower and upper bounds of the support, used in the NPI-B method, of $x_0 = -0.5127$ and $x_{n+1} = 1.7969$.

For sample of size $n = 500$ from the exponential distribution with rate $\lambda = 5$, the sample mean $\bar{x} = 0.1861$. We set the lower and upper bounds of the support, used in the

n_B	$\max(\bar{x}_B)$	$\text{mean}(\bar{x}_B)$	$\min(\bar{x}_B)$	$\hat{RP}(\epsilon)$
100	0.5960	0.2402	0.0918	1.0000
500	0.5960	0.2430	0.0589	1.0000
1000	0.5960	0.2446	0.0097	1.0000
10000	0.7372	0.2496	-0.0512	1.0000
100000	0.9574	0.2499	-0.0672	1.0000

Table 3.5: $\hat{RP}(\epsilon)$ of the mean of characteristics of the exponential distribution with $\lambda = 5$, $n = 30$, $\epsilon = 1$

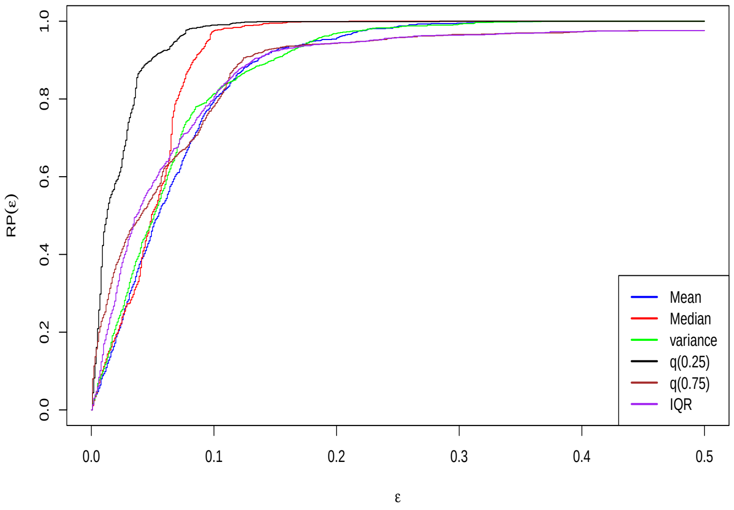


Figure 3.5: $\hat{RP}(\epsilon)$ for estimates of characteristics of the exponential distribution with $\lambda = 5$, $n = 30$, $n_B = 1000$

n_B	$\max(\bar{x}_B)$	$\text{mean}(\bar{x}_B)$	$\min(\bar{x}_B)$	$\hat{RP}(\epsilon)$
100	0.2122	0.1876	0.1553	1.0000
500	0.2259	0.1871	0.1534	1.0000
1000	0.2312	0.187	0.1534	1.0000
10000	0.2375	0.1871	0.1445	1.0000
100000	0.2375	0.1871	0.1445	1.0000

Table 3.6: $\hat{RP}(\epsilon)$ of the mean of characteristics of the exponential distribution with $\lambda = 5$, $n = 500$, $\epsilon = 1$

NPI-B method, of $x_0 = -0.2025$ and $x_{n+1} = 1.4839$ with the maximum distance between consecutive observations $d = 0.1488$.

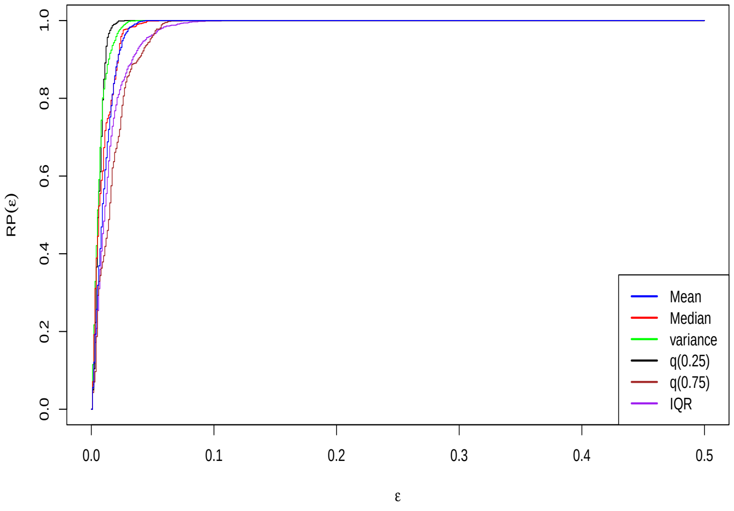


Figure 3.6: $\hat{RP}(\epsilon)$ for estimates of characteristics of the exponential distribution with $\lambda = 5$, $n = 500,\ n_B = 1000$

it is clear from a comparison of Examples 3.5 and 3.6 that in the exponential distribution causes, $\epsilon-$reproducibility of estimates increases earlier than $\epsilon-$reproducibility of estimates of the standard normal distribution. For large sample sizes of $n = 500$, The $\epsilon-$reproducibility of estimates using the exponential distribution closes to 1 at $\epsilon = 0.1$, whereas $\epsilon-$reproducibility of estimates using the standard normal distribution close to 1 at $\epsilon = 0.3$.

It is clear from a comparison of Examples 3.2.1 and 3.2.2, that the exponential distribution causes $\epsilon-$reproducibility of estimates to increase earlier than the standard normal distribution. Most $\epsilon-$reproducibility of estimates values based on the exponential distribution have values close to 1 for $\epsilon = 0.1$, whereas $\epsilon-$reproducibility of estimates values based on the standard normal distribution has values close to 1 for $\epsilon = 0.3$ for large sample sizes of $n = 500$.

3.3 Reproducibility of estimates using a representative sample

This section introduces a new method for assessing $\epsilon-$reproducibility of estimates of a characteristic population using a representative sample procedure instead of NPI-B method. This method helps to avoid randomness of sampling from the distribution. For a population distribution with cumulative distribution function F for real-valued random quantities, we define y_i as a representative sample as follows:

$$y_i = F^{-1}\left(\frac{i}{n+1}\right) \tag{3.4}$$

So, y_i is the $100(\frac{i}{n+1})$-th percentile of F, for $i = 1, ..., n$. We call $Y_1, Y_2, ..., Y_n$ as a original sample of distribution F and order them as $y_{(1)} < y_{(2)} < ... < y_{(n)}$.

The main idea of this method is that we study the reproducibility of estimates of characteristics of F by using the representative sample to give the estimates and to use the NPI method.

As in Section 3.2, we assume finite support in order to simplify the NPI method, so we define the lower and upper bounds of the original sample y_0 and y_{n+1} are derived as follows:

$$y_0 = \min_{1\leq i\leq n}(y_i) - d, \quad y_{n+1} = \max_{1\leq i\leq n}(y_i) + d \tag{3.5}$$

with d again the maximal distance between two consecutive y_i values, where $d = \max_{1\leq i\leq n}(y_i - y_{i-1})$. We now have $n+1$ intervals $I_i = (y_{i-1}, y_i)$ which is determined between the n observations, where $i = 1, ..., n+1$. We assume that all the orderings O_j of the future observations among the original observations are equally likely 1.4, and each ordering includes the future observations $S_i^j = \#\{Y_{n+i}, i = 1, ..., n\}$ where $j = 1, 2, ..., \binom{2n}{n}$. We link the data and future observations via Hill's assumption $A_{(n)}$ [70], or more precisely, via consecutive application of $A_{(n)}, A_{(n+1)}, ..., A_{(2n-1)}$ which can be considered as a post-data version of a finite exchangeability assumption for $2n$ random quantities that are $Y_{n+1}, ..., Y_{2n}$. The $A_{(n)}$ assumptions imply that all possible orderings of n data observations and n future observations are equally likely, where the n data observations and n future observations cannot be separated from one another.

For a larger sample size, we use simple random sampling (SOM)as explained in Section [31] to generate the future observations as explained in Section 1.7.

Based on the $A_{(n)}$ assumptions, Equation (3.6) derive the probability of each ordering [34] as follows.

$$P\left(\bigcap_{i=1}^{n+1}\{S_i^j = s_i^j\}\right) = P(O_j) = \binom{2n}{n}^{-1} \tag{3.6}$$

where the s_i^j are non-negative integers with $\sum_{i=1}^{n} s_i^j = n$.

For ordering O_j, the lower and upper estimates denoted by $\hat{\theta}^f_{j,L}$ and $\hat{\theta}^f_{j,U}$, respectively, can be calculated by using the minimum and maximum possible values the future estimates given these orderings. For example, if interested in the mean, then

$$\hat{\theta}^f_{j,L} = \frac{1}{n}\sum_{i=1}^{n+1} S_i^j y_{i-1}, \qquad \hat{\theta}^f_{j,U} = \frac{1}{n}\sum_{i=1}^{n+1} S_i^j y_i \tag{3.7}$$

We now use these lower and upper estimates corresponding to ordering O_j to derive the lower and upper probabilities for $\epsilon-$reproducibility of the estimates based on a representative sample. This provides a tool to compare RRT as will be explained in examples.

The estimate of $\hat{\theta}^f$ based on the original representative sample is $\hat{\theta}$. To obtain the NPI lower $\epsilon-$reproducibility probability for the event that $|\hat{\theta}^f - \hat{\theta}| \leq \epsilon$, we need to find all estimates $\hat{\theta}^f$ with $[\hat{\theta}^f_{j,L},\ \hat{\theta}^f_{j,U}] \subset [\hat{\theta}-\epsilon, \hat{\theta}+\epsilon]$. To obtain the NPI upper $\epsilon-$reproducibility probability for the event that $|\hat{\theta}^f - \hat{\theta}| \leq \epsilon$ with the condition $[\hat{\theta}^f_{j,L}, \hat{\theta}^f_{j,U}] \cap [\hat{\theta}-\epsilon, \hat{\theta}+\epsilon] \neq \emptyset$ where $j = 1, ..., \binom{2n}{n}$.

This leads to the NPI lower $\epsilon-$reproducibility probability:

$$\underline{RP}(\epsilon) = \underline{P}(|\hat{\theta}^f - \hat{\theta}| \leq \epsilon) = \sum_{j=1}^{\binom{2n}{n}} P(O_j) \ \ \mathbf{1}\left\{ \max(\hat{\theta} - \hat{\theta}^f_{j,L}, \hat{\theta}^f_{j,U} - \hat{\theta}) \leq \epsilon \right\} \tag{3.8}$$

and the NPI upper $\epsilon-$reproducibility probability:

$$\overline{RP}(\epsilon) = \overline{P}(\hat{\theta}^f - \hat{\theta}| \leq \epsilon) = \sum_{j=1}^{\binom{2n}{n}} P(O_j) \ \ \mathbf{1}\left\{ \max(\hat{\theta}^f_{j,L} - \hat{\theta}, \hat{\theta} - \hat{\theta}^f_{j,U}) \leq \epsilon \right\} \tag{3.9}$$

We illustrate this in the following example.

Example 3.3.1

We want to study $\epsilon-$reproducibility of the estimates of characteristics of the standard normal distribution using the representative sample of size $n = 5$ which is $Y_i = \{-0.9674, -0.4307, 0, 0.4307, 0.9674\}$. The lower bound is $y_0 = -1.5041$ and the upper bound is $y_{n+1} = 1.5041$ and $d = 0.5367$. The mean and the median of the original sample are 0, $q(0.75) = -0.4307$, $q(0.75) = 0.4307$ and the IQR is 0.8615.

As can be shown in Figure 3.7, the NPI $\epsilon-$reproducibility of a standard normal distribution with a small sample of size $n = 5$, and the ordering $n_0 = \binom{10}{5} = 252$ of the mean, the median, $q(0.25)$, $q(0.75)$ and the IQR. It is noted that NPI $\epsilon-$reproducibility of median has the lowest maximum values of $\epsilon = \{1.5041, 0.9674, 0.4307\}$ whereas the other IQR have the largest maximum values of $\epsilon = \{0, 0.3248, 0.1059, 0.2119, 0.3248, 0.4307, 0.5367,$
$1.0734,$
$1.61008\}$.

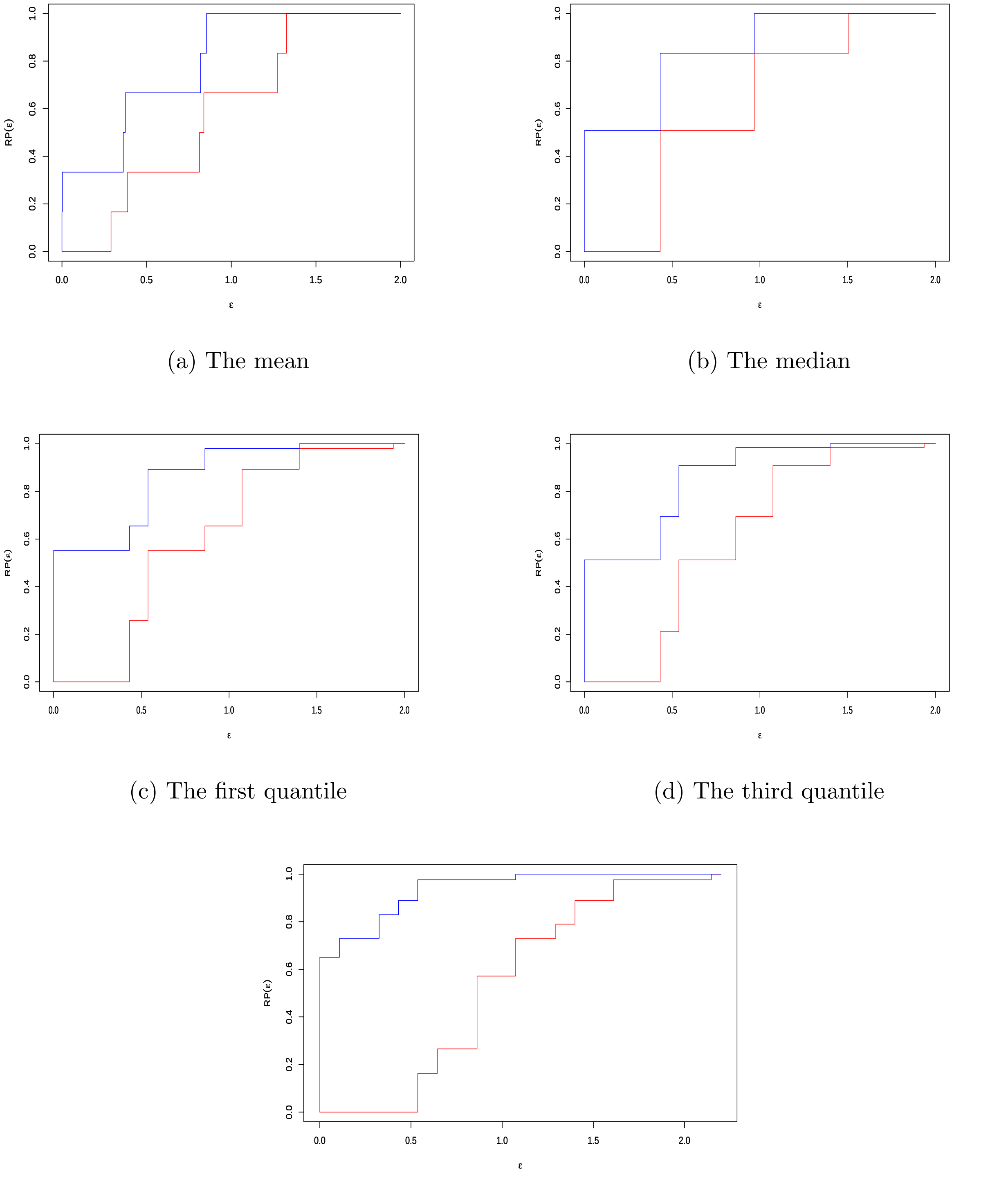
The $\epsilon-$reproducibility of the the mean are 0.6667, 1.0000, the $\epsilon-$reproducibility of the median are 0.8300, 1.0000, the $\epsilon-$reproducibility of the the $q(0.25)$ are 0.6746, 0.9762, the $\epsilon-$reproducibility of the the $q(0.75)$ are 0.6746, 0.9762 and the $\epsilon-$reproducibility of the the IQR are 0.5159, 0.9762.

Note that we use the number of ordering $n_o = \binom{2n}{n}$ where n is a small sample, and it is impossible to derive the $\epsilon-$ reproducibility for a large sample as $n = 500$ and use $n_o = \binom{1000}{500}$. Therefore, we use the sampling of ordering method to solve this issue as explained in Section 1.7.

Example 3.3.2

We want to study $\epsilon-$reproducibility of the estimates of characteristics of the standard normal distribution using the representative sample of size $n = 500$ and the orderings number is large as $n_o = 1000$. We derive the NPI lower $\epsilon-$reproducibility probability using:

$$\underline{RP}(\epsilon) = \underline{P}(|\hat{\theta}^f - \hat{\theta}| \leq \epsilon) = \sum_{j=1}^{n_o} P(O_j) \ \mathbf{1}\left\{ \max(\hat{\theta} - \hat{\theta}^f_{j,L}, \hat{\theta}^f_{j,U} - \hat{\theta}) \leq \epsilon \right\} \tag{3.10}$$



(a) The mean

(b) The median

(c) The first quantile

(d) The third quantile

(e) The IQR

Figure 3.7: The NPI $\epsilon-$reproducibility of standard normal distribution with $n = 5$, $n_0 = \binom{10}{5}$

$n=5$	Mean	Median	$q(0.25)$	$q(0.75)$	IQR
$\hat{RP}(\epsilon)$	0.9880	0.8968	0.9762	0.9722	0.9721
$\underline{RP}(\epsilon)$	0.8929	0.8333	0.6548	0.6944	0.5714
$\overline{RP}(\epsilon)$	1.0000	1.0000	0.9802	0.9841	0.9762

Table 3.7: The $\epsilon-$reproducibility of characteristics of the standard normal distribution of $n_B = n_o = 252$, $\epsilon = 1$

$n=500$	Mean	Median	$q(0.25)$	$q(0.75)$	IQR
$\hat{RP}(\epsilon)$	1.0000	1.0000	0.9880	1.0000	1.0000
$\underline{RP}(\epsilon)$	1.0000	0.9999	0.9980	1.0000	0.9970
$\overline{RP}(\epsilon)$	1.0000	1.0000	1.0000	1.0000	1.0000

Table 3.8: The $\epsilon-$reproducibility of characteristics of the standard normal distribution of $n_B = n_o = 1000$, $\epsilon = 0.3$

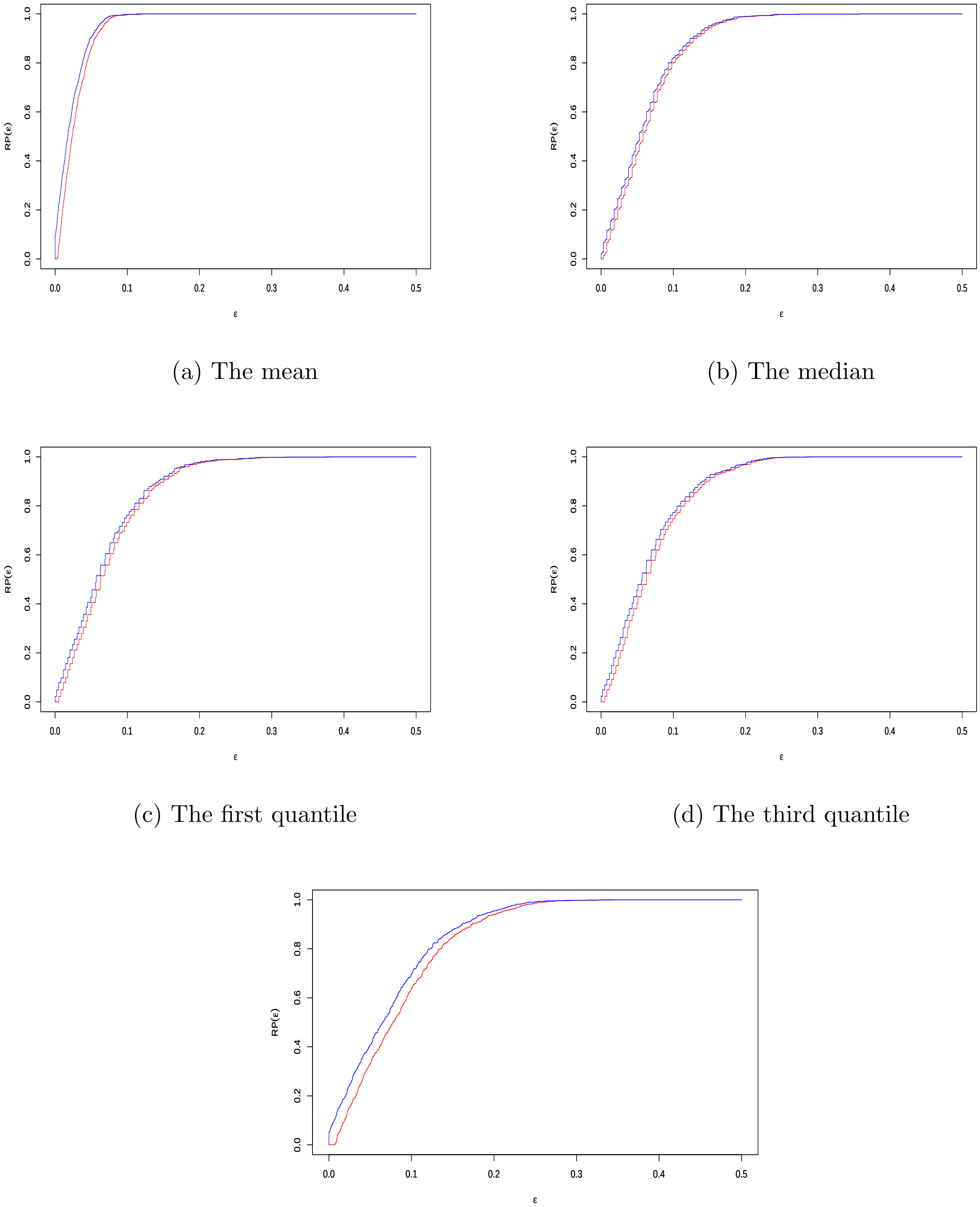
and the NPI upper $\epsilon-$reproducibility probability using

$$\overline{RP}(\epsilon) = \overline{P}(\hat{\theta}^f - \hat{\theta}| \leq \epsilon) = \sum_{j=1}^{n_o} P(O_j) \ \mathbf{1}\left\{ \max(\hat{\theta}^f_{j,L} - \hat{\theta}, \hat{\theta} - \hat{\theta}^f_{j,U}) \leq \epsilon \right\} \tag{3.11}$$

where n_o is the number of orderings and it can be generated using SOM. We illustrate this in the following example.

Figures 3.8 show that increasing the sample size and the ordering numbers leads to better $\epsilon-$reproducibility of estimates. In addition, the $\underline{RP}(\epsilon)$ values closes to $\overline{RP}(\epsilon)$ if the sample size n increases.

Tables 3.7 and 3.8 show the $\epsilon-$reproducibility for estimates using NPI-B method and representative sample. It can be noted that a larger sample size and ordering number or bootstrap numbers lead to higher reproducibility. In this case, we note that the $\epsilon-$reproducibility for estimates using bootstrap have values between the NPI lower and upper $\epsilon-$reproducibility for estimates using the representative sample.



(a) The mean

(b) The median

(c) The first quantile

(d) The third quantile

(e) The IQR

Figure 3.8: The NPI $\epsilon-$reproducibility of standard normal distribution with $n = 500$, $n_o = 1000$, $\epsilon = 0.1$

3.4 Concluding remarks

The method presented in this chapter allows study of reproducibility of estimates either by using NPI-bootstrap or a representative sample.

The NPI- Bootstrap method for quantifying $\epsilon-$reproducibilit of the estimates, as presented in section 3.2, if one has a specific sample and corresponding estimates. However, the main aim of this chapter is to consider reproducibility when different RRT methods are used, without specific samples being available and avoiding the randomness in such samples. In Section 3.3, we introduce a new concept which we call a representative sample of a distribution. The sample we create for the $\epsilon-$reproducibility investigation is not an original sample; rather, it is a way to represent the representative sample using a probability distribution for a population using percentiles of standard normal distribution. Using the representative sample, we do not derive the NPI lower and upper $\epsilon-$reproducibility for the variance. Because the upper variance is a quadratic constraint optimization issue with no typically closed-form solutions. However, in the lower reproducibility for variance, we need to minimise the future sample variance $\hat{\theta}^{f}_{j,L}$ of the m future observations, then derive the probability mass of all observations to the left x_{i-1} or the right x_i in the interval $[x_{i-1}, x_i]$ but we do not know the change point is, it could be good to look for this in the future.

Chapter 4

Reproducibility of estimates based on RRT

4.1 Introduction

In the last chapters, we discussed reproducibility of statistical tests based on RRT data and then we investigate the $\epsilon-$reproducibility of estimates of data generated from the standard normal distribution or the exponential distribution. In this chapter, we use the $\epsilon-$reproducibility methods introduced in Chapter 3 to investigate which RRT methods lead to the best $\epsilon-$reproducibility of estimates using the NPI-B method and the representative sample to compare between RRT methods.

This chapter is structured as follows. The idea of $\epsilon-$reproducibility of estimates based on RRT method using NPI-B method is introduced in Section 4.2. In Section 4.3, the reproducibility of estimates based on a representative sample is investigated. A comparison of the reproducibility for estimates based on RRT data is presented in Section 4.4. Section 4.5 concludes the chapter and presents a consideration of relevant research ideas.

4.2 Reproducibility of the estimates based on RRT methods using NPI-B method

In this section, we derive the $\epsilon-$reproducibility for estimates using data generated by RRT; the approach is detailed in Section 3.2. We use the simulation to generate the original sample of responses of the respondent of different RRT data such as the true response X, the scrambling response S, the response of the unrelated question Y or the reported response Z. These random quantities are generated using a simulation of RRT method, then we use NPI-B method to generate all possible responses which are generated from the original sample. We calculate the original mean $\hat{\mu}_x$ and the future mean $\hat{\mu}_B$ and then we use the $\epsilon-$reproducibility for estimates is explained in Section 3.2. Example 4.2.1 illustrates reproducibility of estimates based on the multiplicative method (MM) method.

Example 4.2.1 This example explains $\epsilon-$reproducibility of the estimate based on real-valued random quantities generated from the multiplicative methods (MM) [62]. Let $X_i \sim N(\mu_x = 4, \sigma_x^2 = 3)$ be the true answer that represents the sensitive characteristic for individual i with an unknown mean μ_x, and let S_i be the scrambling variable. By giving the randomisation device, we generate a random quantity S_i that follows a normal probability distribution with the known mean $E(S_i) = \theta = 1$ and known variance $\gamma^2 = 0.2$.

We assume that all respondents have a probability of scrambling, $\psi = 0.7$. To start with, we simulate a sample with a size of $n = 5$ that represents the reported responses Z_i. The respondent offers the answer $Z_i = X_i$ if the question is not sensitive; if the question is sensitive, the answer is scrambling $Z_i = S_i X_i$.

The simulated values of $X_1,\ X_2,\ X_3,\ X_4,\ X_5$ are 5.0142, 5.2288, 3.8107, 3.2145, 5.0494, and the simulated values of $S_1,\ S_2,\ S_3,\ S_4,\ S_5$ are 0.1870, 1.2818, 0.8765, 0.8729, 0.5889, this leads to $X_i S_i$ takes values 0.9376, 6.7023, 3.3400, 2.8060, 2.9734. Assume that the randomisation device generated the values $1, 1, 0, 0, 1$. If the value is 1, the response is $Z_i = X_i S_i$. If 0, the answer is $Z_i = X_i$. So, the actual values are $Z_1 = 0.9376, Z_2 = 6.7023, Z_3 = 3.8107, Z_4 = 3.2145$ and $Z_5 = 2.9734$ and the sample mean is

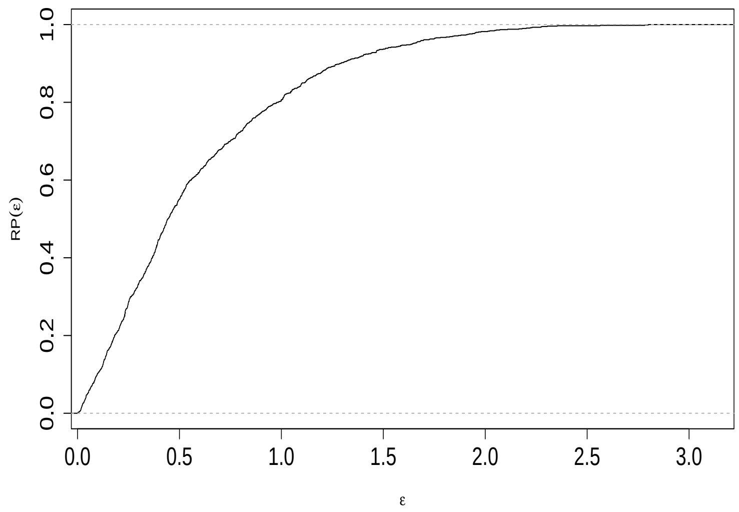


Figure 4.1: The $\hat{RP}(\epsilon)$ of the MM, $n = 5,\ n_B = 1000,\ \mu_x = 4,\ \sigma_x^2 = 3,\ \theta = 1,\ \gamma^2 = 0.2,$ $\psi = 0.70$

$\mu_z = 3.3300$.

As discussed in Section 3.2, in order to implement NPI-B method to assess the $\epsilon-$reproducibility of estimates, we calculate the bounds of support of Z_i as follows.

$$z_0 = \min_{1 \leq i \leq n}(z_i) - d = -0.2381, \qquad z_{n+1} = \max_{1 \leq i \leq n}(z_i) + d = 7.4675 \tag{4.1}$$

where $d = 2.0526$ is the maximal distance between two consecutive z_i values.

We generate $n_B = 1000$ bootstrapping samples each consisting 5 values and we calculate the bootstrap estimate for the sample mean $\hat{\mu}_{B_i}$, for each bootstrap sample i. The $\epsilon-$reproducibility of the estimate of the mean is

$$\hat{RP}(\epsilon) = \sum_{i=1}^{n_B} \frac{1}{n_B} \mathbf{1}\Big\{|\hat{\mu}_x - \hat{\mu}_{B_i}| \leq \epsilon\Big\}, \qquad \text{where} \quad \epsilon \geq 0 \tag{4.2}$$

Figure 4.1 shows that $\hat{RP}(\epsilon)$ is a function of ϵ, where $\epsilon \in [0, 3]$. The results illustrate clearly that the lowest value of $\hat{RP}(\epsilon)$ is for $\epsilon = 0$ whereas the highest value of $\hat{RP}(\epsilon)$ for $\epsilon = 3$. The increasing of ϵ leads to increasing of $\hat{RP}(\epsilon)$. Therefore, for any two values $\epsilon_2 > \epsilon_1$, the $\hat{RP}(\epsilon_2) > \hat{RP}(\epsilon_1)$.

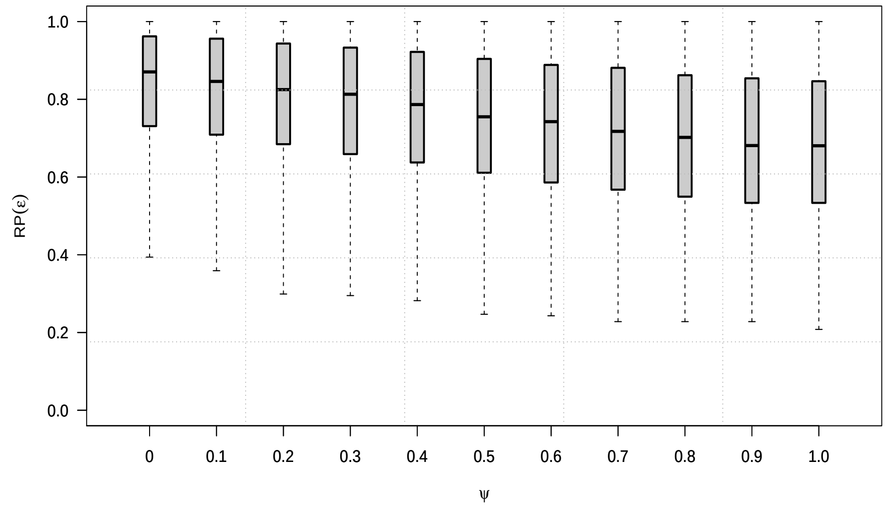


Figure 4.2: $\hat{RP}(\epsilon)$ versus ψ based on the MM with $n = 5$, $n_B = 1000$, $\mu_x = 4$, $\sigma_x^2 = 3$, $\theta = 1$, $\gamma^2 = 0.2$, $\epsilon = 1$

Summary	$\psi = 0$	$\psi = 0.1$	$\psi = 0.2$	$\psi = 0.3$	$\psi = 0.4$	$\psi = 0.5$	$\psi = 0.6$	$\psi = 0.7$	$\psi = 0.8$	$\psi = 0.9$	$\psi = 1.0$
$q(0.25)$	0.7310	0.709	0.68475	0.6590	0.6378	0.6115	0.5865	0.5678	0.5498	0.5338	0.5338
$q(0.75)$	0.9620	0.9560	0.9433	0.9330	0.9220	0.9040	0.8883	0.8810	0.8615	0.8540	0.8463
median	0.8705	0.8460	0.8250	0.8130	0.7865	0.7550	0.7425	0.7175	0.7020	0.6810	0.6805
mean	0.8374	0.8193	0.8019	0.7858	0.7681	0.7476	0.7288	0.7134	0.6989	0.6857	0.6833
sd	0.1407	0.1498	0.15688	0.1649	0.1715	0.1786	0.1810	0.1858	0.1894	0.1926	0.1954
IQR	0.2310	0.2470	0.2585	0.2740	0.2843	0.2925	0.3018	0.3133	0.3118	0.3203	0.3125
lowest whisker	0.3845	0.3385	0.2970	0.2480	0.2114	0.1728	0.1339	0.0979	0.0821	0.0534	0.0650
highest whisker	1	1	1	1	1	1	1	1	1	1	1

Table 4.1: $\hat{RP}(\epsilon)$ of estimates of the MM with $n = 5$, $n_B = 1000$, $\mu_x = 0.2$, $\sigma_x^2 = 3$, $\theta = 1$, $\gamma^2 = 0.2$, $\epsilon = 1$

For many replications $n^* = 100$, we generate different original samples and derive the $\hat{RP}(1)$ versus ψ as shown in Figure 4.2. It is noted that the reproducibility tends to decrease as the ψ increases. At $\epsilon = 1$, the respondents do not use the scrambling variables which means the question is not sensitive. So, a large proportion of people are not scrambling which leads to less variation between the responses.

To estimate the summary of $\hat{RP}(\epsilon)$ of estimates of different ψ, Table 4.1 shows that the $\hat{RP}(\epsilon)$ of the medians shows a decreasing if ψ increases, from 0.87 to 0.68. The $\hat{RP}(\epsilon)$ of the means are always less than the medians except for the means of $\hat{RP}(\epsilon)$ of $\psi = \{0.9, 1\}$.

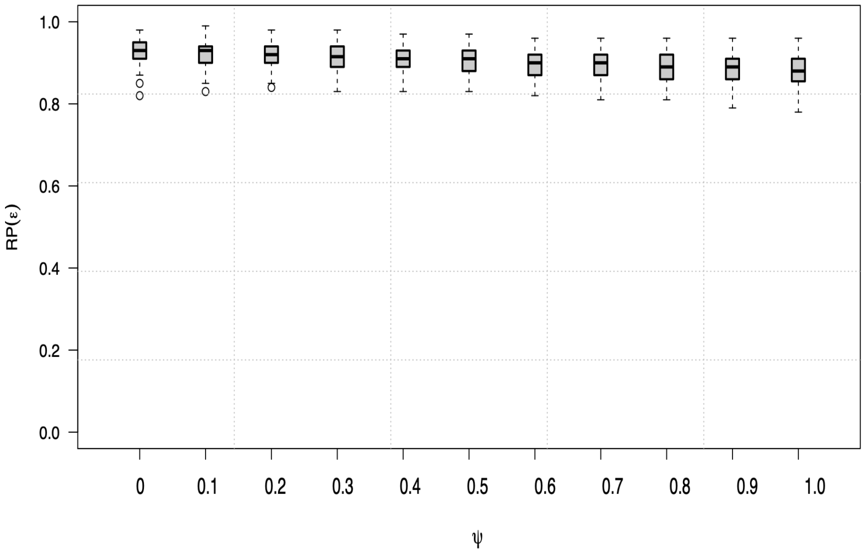


Figure 4.3: $\hat{RP}(\epsilon)$ of the MM of $n = 500,\ n_B = 100, n^* = 100, \mu_x = 4, \sigma_x^2 = 20,\ \theta = 1,$ $\gamma^2 = 0.2,\ \epsilon = 0.5$

Summary	$\psi = 0$	$\psi = 0.1$	$\psi = 0.2$	$\psi = 0.3$	$\psi = 0.4$	$\psi = 0.5$	$\psi = 0.6$	$\psi = 0.7$	$\psi = 0.8$	$\psi = 0.9$	$\psi = 1.0$
$q(0.25)$	0.91	0.9	0.9	0.89	0.89	0.88	0.87	0.87	0.86	0.86	0.8575
$q(0.75)$	0.95	0.94	0.94	0.94	0.93	0.93	0.92	0.92	0.92	0.91	0.91
median	0.93	0.93	0.92	0.915	0.91	0.91	0.9	0.9	0.89	0.89	0.88
mean	0.9267	0.9232	0.9193	0.9145	0.9095	0.9055	0.8987	0.8934	0.8901	0.8849	0.8785
sd	0.0295	0.0288	0.0289	0.0302	0.0301	0.0311	0.0336	0.0351	0.0352	0.0366	0.0398
IQR	0.04	0.04	0.04	0.05	0.04	0.05	0.05	0.05	0.06	0.05	0.053
lowest whisker	0.85	0.84	0.84	0.815	0.83	0.805	0.795	0.795	0.770	785	0.779
highest whisker	1	1	1	1	0.990	1	0.995	0.995	1	0.9850	0.9888

Table 4.2: $\hat{RP}(\epsilon)$ for different value of ψ of the MM of $n = 500,\ n_B = 100,$ $n^* = 100,\ \mu_x = 4, \sigma_x^2 = 20,\ \theta = 1,\ \gamma^2 = 0.2,\ \epsilon = 0.5$

The $\hat{RP}(\epsilon)$ of IQR takes values between 0.23 and 0.31. In addition, the $\hat{RP}(\epsilon)$ of $q(0.25)$ takes value between 0.73 and 0.53 whereas the $\hat{RP}(\epsilon)$ of $q(0.75)$ takes values between 0.96 and 0.84. Conversely, the $\hat{RP}(\epsilon)$ of standard deviation increase if ψ increases. The lowest whisker stakes value between 0.06 and 0.38 whereas the highest whiskers take values exceeds 1.

When the sample size is increased to $n = 500$, the reproducibility increases because the difference between the estimates of the original sample and the future sample decrease which leads to increasing the $\hat{RP}(1) = 1$ of all characteristics except sd and IQR which are equal to 0.

The length of the rectangles in the boxplot shows the variations in $\epsilon-$reproducibility. It is noted that $\epsilon-$reproducibility decreases if ϵ decreases of all values of scrambling ψ.

Large sample size n leads to higher reproducibility because the difference between the original estimate and the future estimate decreases, but large replications n^* do not lead to more changes in the reproducibility because larger replication n^* leads to more accurate reproducibility for estimates.

Example 4.2.2 This example explains $\epsilon-$reproducibility of the estimate based on real-valued random quantities generated from the multiplicative methods (MM) [62]. Let $X_i \sim N(\mu_x = 4,\ \sigma_x^2 = 20)$ be the true answer that represents the sensitive characteristic for individual i with an unknown mean μ_x, and let S_i be the scrambling variable. By giving the randomisation device, we generate a random quantity S_i that follows a normal probability distribution with the known mean $E(S_i) = \theta = 1$ and known variance $\gamma^2 = 10$.

To start with, we simulate a sample with a size of $n = 5$ that represents the reported responses Z_i. In this example, we increase the variance of normal distribution for each X_i and S_i. For $\sigma_x^2 = 20$, it is noted that $\hat{RP}(0.5)$ decreases if ψ decreases. The largest reproducibility is for $\psi = 0$ whereas the lowest reproducibility is for $\psi = 1$. Figure 4.3 and Table 4.2 show that reproducibility decreases slightly if the sensitivity level ψ decreases.

The $\hat{RP}(0.5)$ of the mean takes values between 0.92 and 0.87 and the $\hat{RP}(0.5)$ of the median takes value between 0.93 and 0.88. The $\hat{RP}(0.5)$ of $q(0.25)$ takes values between 0.91 and 0.85 whereas the $\hat{RP}(0.5)$ of $q(0.75)$ takes values between 0.95 and 0.91. The $\hat{RP}(0.5)$ of the IQR takes values between 0.04 and 0.06. The $\hat{RP}(0.5)$ of the standard deviation sd takes small values between 0.02 and 0.03. The lowest whisker takes values between 0.85 and 0.77 whereas the highest whisker takes values between 0.90 and 0.98.

Now, we increase the variance of normal distribution of S_i as shown in Figure 4.4 and Table 4.3. It is noted that the $\epsilon-$reproducibility of estimates for all characteristics at $\gamma^2 = 30$ decrease clearly if ψ increases more than the $\epsilon-$reproducibility of estimates at

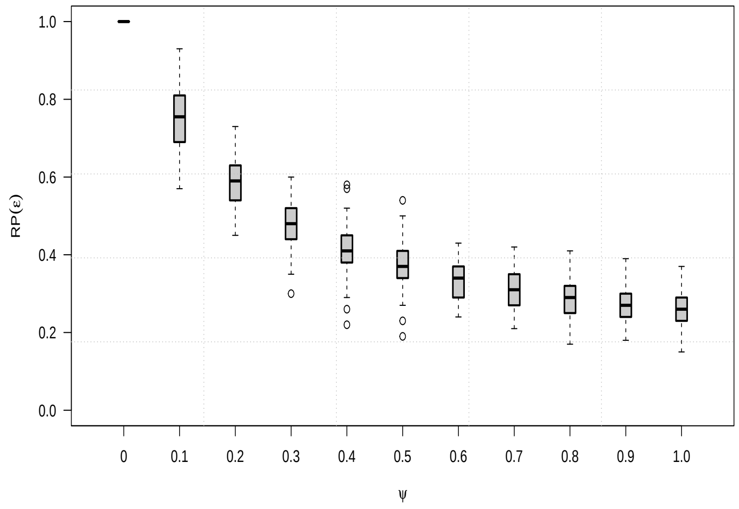


Figure 4.4: $\hat{RP}(\epsilon)$ based on the MM, $n = 500,\ n_B = 100,\ n^* = 100,\ \mu_x = 4,\ \theta = 1,\ \sigma_x^2 = 3,\ \gamma^2 = 30,\ \epsilon = 0.5$

Summary	$\psi = 0$	$\psi = 0.1$	$\psi = 0.2$	$\psi = 0.3$	$\psi = 0.4$	$psi = 0.5$	$\psi = 0.6$	$\psi = 0.7$	$\psi = 0.8$	$\psi = 0.9$	$\psi = 1.0$
$q(0.25)$	1	0.6900	0.5400	0.4400	0.3800	0.3400	0.2900	0.2700	0.2500	0.2400	0.2300
$q(0.75)$	1	0.8100	0.6300	0.5200	0.4500	0.4100	0.3700	0.3500	0.3200	0.3000	0.2900
median	1	0.7550	0.5900	0.4800	0.4100	0.3700	0.3400	0.3100	0.2900	0.2700	0.2600
mean	1	0.7499	0.5827	0.4800	0.4125	0.3733	0.3357	0.3121	0.29	0.273	0.2606
sd	0	0.0802	0.0643	0.0578	0.0582	0.0565	0.0488	0.0482	0.0456	0.0419	0.0428
IQR	0	0.1200	0.0900	0.0800	0.0700	0.0700	0.0800	0.0800	0.0700	0.0600	0.0600
lowest whisker	1	0.5100	0.4050	0.3200	0.2750	0.2350	0.1700	0.1500	0.1450	0.1500	0.1400
highest whisker	1	0.9900	0.7650	0.6400	0.5550	0.5150	0.4900	0.4700	0.4250	0.3900	0.3800

Table 4.3: $\hat{RP}(\epsilon)$ estimates of the MM of $n = 500,\ n_B = 100,\ n^* = 100,\ \mu_x = 4,\ \theta = 1,\ \sigma_x^2 = 3,\ \gamma^2 = 30,\ \epsilon = 0.5$

$\gamma^2 = 0.2$. Therefore, the increased variance of the normal distribution of S_i leads to poor reproducibility, even if the sample size increases.

Similarly, we introduce examples to the $\epsilon-$reproducibility of the estimate based on real-valued random quantities generated from the Greenberg method (GM) as explained in Section 1.2.1.

Example 4.2.3 This example introduces $\epsilon-$reproducibility of the estimate based on real-valued random quantities generated from the Greenberg method [61]. Respondents use the randomisation device to answer one of two questions. One of these questions is sensitive while the other is nonsensitive. Both answers are real-valued quantities.

Assume the probability of the sensitive question is $\gamma = 0.70$. We simulate the responses to the sensitive question $X_i \sim N(\mu_x = 1, \sigma_x^2 = 10)$ and the responses to the unrelated question $Y_i \sim N(\mu_y = 4, \sigma_y^2 = 20)$.

Suppose that for $n = 5$, we represent the X_i values $2.8516, 3.2435, 0.6544$, -0.4341, 2.9160 and the Y_i values are are $-4.1301, 6.8179, 2.7649$, $2.7292, -0.1113$. These values are simulated from the given distributions.

Assume that $\gamma = 0.70$ is the probability of the question of interest being X_i for each person in which can they give Y_i as an answer. Assume that the randomisation device generated the values $\{1, 1, 0, 0, 1\}$. If the value 1, the response is $Z_i = Y_i$. If 0, the answer is $Z_i = X_i$.

The reported Z_i responses are $-4.1301, 6.8179, 0.6544, -0.4341, -0.1113$. The estimate of the reported responses Z_i based on the response X_i to the sensitive question $\hat{\mu}_x^z$ is

$$\hat{\mu}_x^z = \frac{\hat{\mu}_z - (1-\gamma)\mu_y}{\gamma} = -0.9152 \tag{4.3}$$

where $\hat{\mu}_z = \frac{\sum_{i=1}^{n} Z_i}{n} = 0.5593$.

To apply NPI-B method for determining the lower and upper bounds for the support Z_i:

$$z_0 = \min_{1 \le i \le n} (z_i) - d = -10.2937 \tag{4.4}$$

$$z_{n+1} = \max_{1 \le i \le n} (z_i) + d = 12.9814 \tag{4.5}$$

where $d = 6.1635$ is the maximal distance between two consecutive z_i values. We generate $n_B = 1000$ NPI-B samples $b_1, ...b_n$ size n based on the z_i values.

We calculate the expected value of bootstrap samples b_i based on the sample Z_i as $\hat{\mu}_z^B$, and we use the mean of normal distribution of the unrelated responses μ_y to derive the estimate of each bootstrap sample $\hat{\mu}_x^B$ as follows:

$$\hat{\mu}_x^B = \frac{\hat{\mu}_z^B - (1-\gamma)\mu_y}{\gamma} \tag{4.6}$$

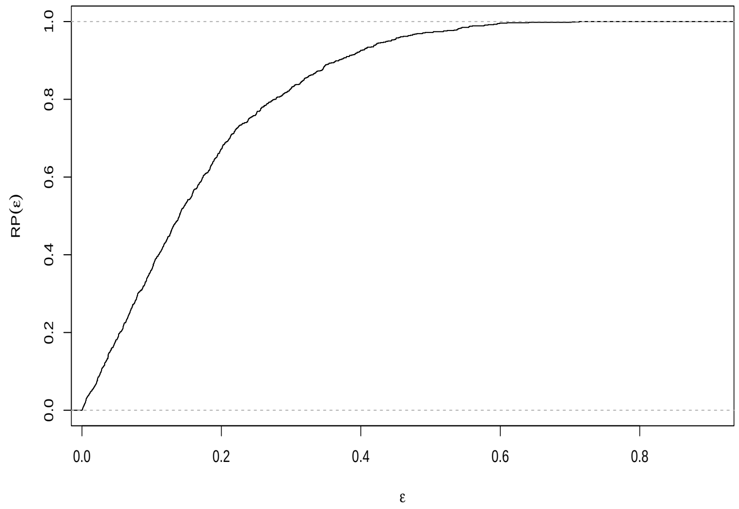


Figure 4.5: The average of $\hat{RP}(\epsilon)$ of the Greenberg method, $n = 5,\ n_B = 100,\ n^* = 1000,$ $\mu_x = 1,\ \sigma_x^2 = 10,\ \mu_y = 4,\ \sigma_y^2 = 20, \gamma = 0.70$

where the mean of each bootstrap sample is $\hat{\mu}_z^B = \frac{\sum_{i=1}^n b_i}{n}$. Then, we calculate the difference between $\hat{\mu}_x^z$ and bootstrap samples means $\hat{\mu}_x^B$ to derive $\epsilon -$ reproducibility of the mean for n_B times. Then, find the number of of the event that $|\hat{\mu}_x^z - \hat{\mu}_x^B| \leq \epsilon$ divided by n_B, as follows:

$$|\hat{\mu}_x^z - \hat{\mu}_x^B| \leq \epsilon \Longleftrightarrow \left| \frac{\hat{\mu}_z - (1-\gamma)\mu_y}{\gamma} - \frac{\hat{\mu}_z^B - (1-\gamma)\mu_y}{\gamma} \right| \leq \epsilon \tag{4.7}$$

This leads to derive $\epsilon -$ reproducibility of the mean as follows:

$$\hat{RP}(\epsilon) = P\bigg(|\hat{\mu}_x^z - \hat{\mu}_x^B| \leq \epsilon \bigg) = \sum_{i=1}^{n_B} \frac{1}{n_B}\ \mathbf{1}\left\{ \left| \frac{\hat{\mu}_z}{\gamma} - \frac{\hat{\mu}_x^{B_i}}{\gamma} \right| \leq \epsilon \right\} \tag{4.8}$$

Perform this procedure n^* times to get n^* original sample to derive n^* of $\epsilon-$reproducibility for estimates.

Figure 4.5 shows the $\hat{RP}(\epsilon)$ as function. It can be seen that for larger values of ϵ, reproducibility increases, because the difference between the estimate based on the original sample and the estimate based on the future sample is small which increases the $\hat{RP}(\epsilon)$. The results illustrate clearly that the lowest value of $\hat{RP}(\epsilon)$ is for $\epsilon = 0$ whereas the highest value of $\hat{RP}(\epsilon)$ for $\epsilon = 3$. The increasing of ϵ leads to increasing of $\hat{RP}(\epsilon)$. Therefore, for any two values $\epsilon_2 > \epsilon_1$, the $\hat{RP}(\epsilon_2) > \hat{RP}(\epsilon_1)$.

Summary	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$	$\gamma = 1.0$
$q(0.25)$	0.0270	0.0530	0.0760	0.0978	0.1160	0.1360	0.1560	0.1740	0.1950	0.3008
$q(0.75)$	0.0500	0.0940	0.1340	0.1690	0.2040	0.2283	0.2613	0.2965	0.337	0.4960
median	0.0370	0.0690	0.1010	0.1280	0.1500	0.1710	0.1950	0.2215	0.2470	0.3830
mean	0.0424	0.0797	0.1155	0.1435	0.1711	0.1957	0.2240	0.2539	0.2820	0.4142
sd	0.0218	0.0419	0.0657	0.0698	0.0815	0.0970	0.1071	0.1192	0.1271	0.1549
IQR	0.0230	0.0410	0.0580	0.0713	0.0880	0.0923	0.1053	0.1225	0.1420	0.19525
lowest whisker	0	0	0	0	0	0	0	0	0	0
highest whisker	0.0845	0.1555	0.2210	0.2759	0.3360	0.3666	0.4191	0.4803	0.5500	0.7889

Table 4.4: $\hat{RP}(\epsilon)$ of the GM method of $n = 5, n_B = 100,\ n^* = 100, \mu_x = 1,\ \sigma_x^2 = 10,$ $\mu_y = 4,\ \sigma_y^2 = 20,\ \epsilon = 1$

For different original samples, the reproducibility for estimates based on Greenberg's method can be visualised using a boxplot with different $\gamma \in [0, 1]$ where the variances of the normal distributions $\sigma_x^2 > \sigma_y^2$.

Figure 4.6 shows that reproducibility gets higher as γ becomes larger because many people answer the unrelated question. The $\hat{RP}(1)$ of the mean takes values between 0.04 and 0.28 and the $\hat{RP}(1)$ of the median takes value between 0.03 and 0.27. The $\hat{RP}(1)$ of $q(0.25)$ takes values between 0.02 and 0.19 whereas the $\hat{RP}(1)$ of $q(0.75)$ takes values between 0.05 and 0.33. The $\hat{RP}(1)$ of the IQR takes values between 0.02 and 0.14. The $\hat{RP}(1)$ of the standard deviation sd takes small values between 0.02 and 0.12. The highest whisker takes values between 0.08 and 0.55.

A large sample size leads to higher reproducibility. However, a large number of replications n^* does not lead to more changes in the reproducibility because larger replication leads to more accurate for reproducibility for estimates as shown in Figure 4.5 and Table 4.5.

For the assumptions, $\mu_y > \mu_x$ and $\sigma_y^2 > \sigma_x^2$, an increasing the variance of the distribution of the non-sensitive answers leads to an increase of the $\epsilon-$reproducibility for the estimates as shown in Figures 4.7 and 4.8 respectively. Tables 4.5 shows the $\epsilon-$reproducibility of estimates are higher than the $\epsilon-$reproducibility of estimates in Table 4.6 shows.

The $\hat{RP}(0.5)$ of the mean takes values between 0.17 and 0.98 and the $\hat{RP}(0.5)$ of the median takes value between 0.17 and 0.90. The $\hat{RP}(0.5)$ of $q(0.25)$ takes values between

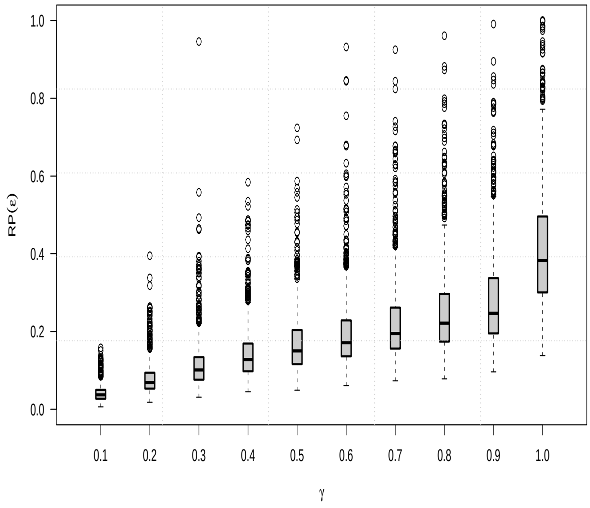


Figure 4.6: $\hat{RP}(\epsilon)$ of the GM method of $n = 5,\ n_B = 100, n^* = 100,\ \mu_x = 1,\ \sigma_x^2 = 10,$ $\mu_y = 4,\ \sigma_y^2 = 20,\ \epsilon = 1$

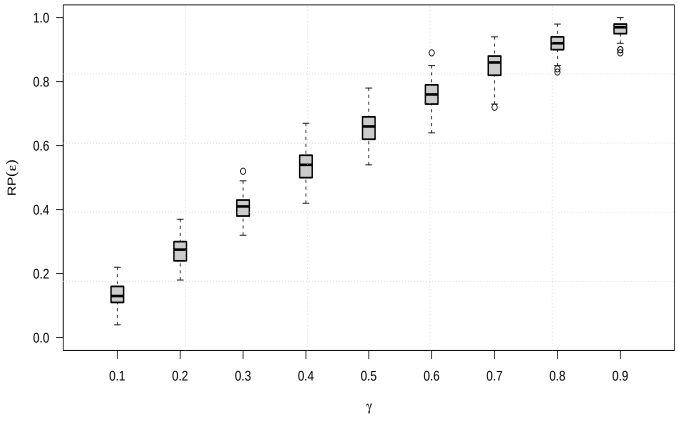


Figure 4.7: $\hat{RP}(\epsilon)$ of the GM method of $n = 500, n_B = 100, n^* = 100, \mu_x = 1,\ \sigma_x^2 = 10,$ $\mu_y = 4,\ \sigma_y^2 = 20,\ \epsilon = 0.5$

0.15 and 0.87 whereas the $\hat{RP}(0.5)$ of $q(0.75)$ takes values between 0.2 and 0.91. The $\hat{RP}(0.5)$ of the IQR takes values between 0.05 and 0.07. The $\hat{RP}(0.5)$ of the standard deviation sd takes small values between 0.03 and 0.05. The highest whisker takes values between 0.07 and 0.81. and the lowest whisker takes values between 0.27 and 0.97.

Now, we increase the mean of the distribution of the sensitive responses and investigate

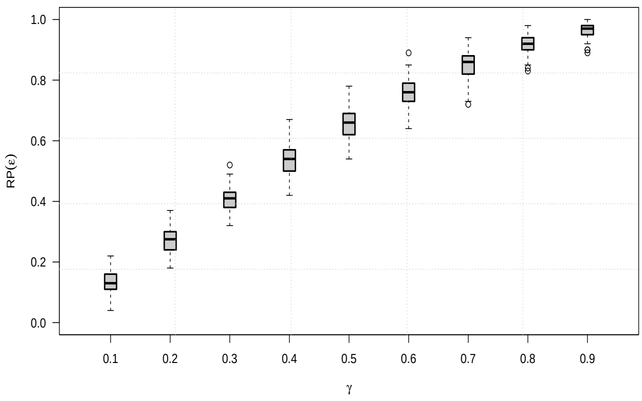


Figure 4.8: $\hat{RP}(\epsilon)$ of the GM method of $n = 500,\ n_B = 100,\ n^* = 100,\ \mu_x = 1, \sigma_x^2 = 20,$ $\mu_y = 4,\ \sigma_y^2 = 10,\ \epsilon = 0.5$

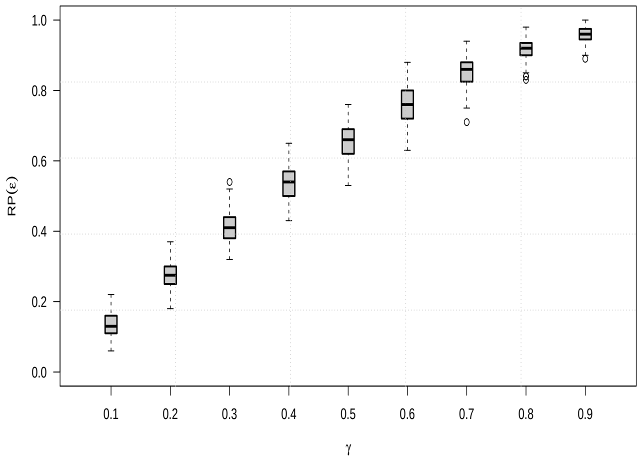


Figure 4.9: $\hat{RP}(\epsilon)$ of the GM method, $n = 500,\ n_B = 100,\ n^* = 100,\ \mu_x = 4, \sigma_x^2 = 20,$ $\mu_y = 1,\ \sigma_y^2 = 10,\ \epsilon = 0.5$

the $\epsilon-$reproducibility of estimates where $\mu_x > \mu_y$ and $\sigma_y^2 < \sigma_x^2$. We find that the $\epsilon-$reproducibility of estimates of all characteristics increases than the $\epsilon-$reproducibility of estimates at $\mu_x < \mu_y$ and $\sigma_y^2 < \sigma_x^2$ except the $\epsilon-$reproducibility of the median as shown in Figure 4.9 and Table 4.7.

So, the $\epsilon-$reproducibility of estimates of the GM gets higher values if the mean of the distribution of the sensitive responses and the variance of the distribution of the unrelated

Summary	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
$q(0.25)$	0.1500	0.3000	0.4275	0.5300	0.6300	0.7000	0.7700	0.8300	0.8700
$q(0.75)$	0.2000	0.3600	0.4825	0.6000	0.6900	0.7600	0.8200	0.8800	0.9100
median	0.1700	0.3300	0.4600	0.5600	0.6500	0.7300	0.8000	0.8500	0.900
mean	0.1736	0.3301	0.4578	0.5645	0.6551	0.7309	0.7967	0.8496	0.8902
sd	0.0358	0.0438	0.0440	0.0531	0.0506	0.0446	0.0462	0.04144	0.0335
IQR	0.0500	0.0600	0.0550	0.0700	0.0600	0.0600	0.0500	0.0500	0.0400
lowest whisker	0.0750	0.2100	0.3450	0.4250	0.5400	0.6100	0.6950	0.7550	0.8100
highest whisker	0.2750	0.4500	0.5650	0.7050	0.7800	0.8500	0.8950	0.9550	0.9700

Table 4.5: $\hat{RP}(\epsilon)$ of the GM method of $n = 500, n_B = 100,\ n^* = 100,\ \mu_x = 1, \sigma_x^2 = 10,$ $\mu_y = 4,\ \sigma_y^2 = 20,\ \epsilon = 0.5$

Summary	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
$q(0.25)$	0.1100	0.2400	0.3800	0.500	0.6200	0.7300	0.8200	0.900	0.9500
$q(0.75)$	0.1600	0.3000	0.4300	0.5700	0.6900	0.7900	0.8800	0.9400	0.9800
median	0.1300	0.2750	0.4100	0.5400	0.6600	0.7600	0.8600	0.9200	0.9700
mean	0.1345	0.2736	0.4077	0.5385	0.6553	0.7595	0.8498	0.9177	0.9602
sd	0.0343	0.0408	0.0425	0.0526	0.0539	0.0449	0.04192	0.0307	0.0231
IQR	0.0500	0.0600	0.0500	0.0700	0.0700	0.0600	0.0600	0.0400	0.0300
lowest whisker	0.0350	0.1500	0.3050	0.3950	0.5150	0.6400	0.7300	0.8400	0.9050
highest whisker	0.2350	0.3900	0.5050	0.6750	0.7950	0.8800	0.9700	1.0000	1

Table 4.6: $\hat{RP}(\epsilon)$ of the GM method of $n = 500,\ n_B = 100,\ n^* = 100, \mu_x = 1, \sigma_x^2 = 20,$ $\mu_y = 4,\ \sigma_y^2 = 10,\ \epsilon = 0.5$

Summary	$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.9$
$q(0.25)$	0.1600	0.3000	0.4400	0.5700	0.6900	0.8000	0.8800	0.9325	0.9725
$q(0.75)$	0.1600	0.3000	0.4400	0.5700	0.6900	0.8000	0.8800	0.9325	0.9725
median	0.1300	0.2750	0.4100	0.5400	0.6600	0.7600	0.8600	0.9200	0.9600
mean	0.1359	0.2742	0.4102	0.5403	0.6569	0.7616	0.8516	0.9163	0.9593
sd	0.0344	0.0405	0.0445	0.0502	0.0506	0.0493	0.0449	0.0305	0.0219
IQR	0.0500	0.0500	0.0600	0.0700	0.0700	0.0800	0.0525	0.0325	0.0250
lowest whisker	0.0350	0.1750	0.2900	0.3950	0.5150	0.6000	0.7488	0.8513	0.9100
highest whisker	0.2350	0.3750	0.5300	0.6750	0.7950	0.9200	0.9588	0.98125	1

Table 4.7: $\hat{RP}(\epsilon)$ of the GM method of $n = 500,\ n_B = 100,\ n^* = 100, \mu_x = 4, \sigma_x^2 = 20,$ $\mu_y = 1,\ \sigma_y^2 = 10,\ \epsilon = 0.5$

response increase.

In general, as γ increases, the $\epsilon-$reproducibility of the estimate of the GM rises. It is to be noted that while the $\epsilon-$reproducibility of estimates of the MM gets higher values than the $\epsilon-$reproducibility of estimates of the GM with large sample size, the $\epsilon-$reproducibility of estimates of the GM has less variation with small sample size than the $\epsilon-$reproducibility of estimates of the MM.

Both the $\epsilon-$reproducibility of estimates of MM and GM are investigated using a single sample. So, it is useful now to investigate the process of deriving the additive method's estimates' $\epsilon-$reproducibility using two samples.

Example 4.2.4

This example explains reproducibility for estimates based on the additive method (AM) as explained in Section 1.2.2. Assume that a sample of size n is divided into two sub-samples of sizes $n_1 = 6$ and $n_2 = 4$, where $n_1 + n_2 = n$. Let assume true response X be simulated from $N(\mu_x = 4, \sigma_x^2 = 3)$. Let $S_1 \sim N(\theta_1 = 3, \sigma_{s_1}^2 = 2)$ and $S_2 \sim N(\theta_2 = 5, \sigma_{s_2}^2 = 4)$. where $\theta_1 \neq \theta_2$. Suppose the sensitivity level is $\psi = 0.70$ which is known. Therefore, the simulated true responses X_i are: $5.0142, 5.2288, 3.8107, 3.2145, 5.0494,$ $0.8512, 5.0914, 3.5216, 3.5078, 2.4077$. The scrambling responses are: S_1 are $2.8356, 5.5701,$ $3.5241, 3.7357, 1.9386, 4.1553$ and S_2 are $6.5592, 7.9116, 3.7113, 5.5974$. Each individual i is asked one of two questions using the randomisation device. The level of sensitivity of the sensitive question is $\psi = 0.7$, and the scrambled response is $X + S_j$, where $j = 1, 2$. Draw 0 and 1 variables in a sample V_j such that $V_j = 1$ has probability ψ and $V_j = 0$ otherwise, as follows:
$V_1 = (1, 1, 1, 1, 0, 0)$ that means the responses are $Z_1 = \{X_1 + S_1, X_2 + S_1, X_3 + S_3, X_4 + S_4,$ $X_5, X_6\}$, and $V_2 = (1, 1, 1, 1)$ that means the responses are $Z_2 = \{X_7 + S_7, X_8 + S_8, X_9 +$ $S_9, X_{10} + S_{10}\}$.
Then, the reported responses Z_j are: $X_1 + S_1 = 7.8497$, $X_2 + S_2 = 10.7989$, $X_3 =$ 7.3348, $X_4 = 6.9502$,and $X_5 = 5.0494$, $X_6 = 0.8512$, $X_7 + S_7 = 10.6889$, $X_8 + S_8 =$ 10.0808, $X_9 + S_9 = 11.4193$ and $X_{10} + S_{10} = 6.1190$. The expected values of Z_1 and Z_2 are

$$E(Z_1) = \mu_x + \psi\theta_1 = 6.1 \tag{4.9}$$

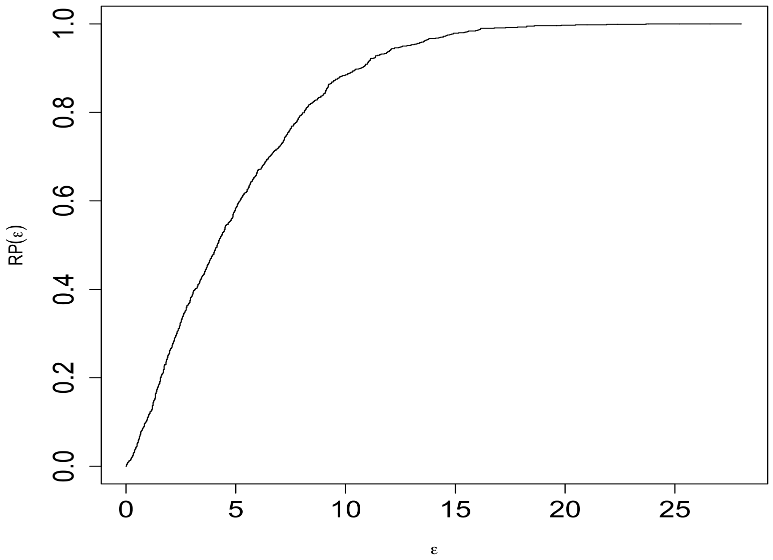


Figure 4.10: $\hat{RP}(\epsilon)$ of the AM method, $n = 10, n_1 = 6,$ n$_2$ $= 4,\ n_B = 1000, n^* = 1000, \mu_x = 4,\ \sigma_x^2 = 3,\ \theta_1 = 3,\ \theta_2 = 5,\ \sigma_{s_1}^2 = 2,\ \sigma_{s_2}^2 = 4,\ \psi = 0.7$

$$E(Z_2) = \mu_x + \psi\theta_2 = 7.5 \tag{4.10}$$

To apply NPI-B method, we need to choose the lower and upper bounds for the support Z_j:

$$z_0 = \min_{1 \le i \le n} (z_j) - d \tag{4.11}$$

$$z_{n+1} = \max_{1 \le i \le n} (z_j) + d \tag{4.12}$$

where d is the maximal distance between two consecutive z_j values. Therefore, the lower and upper bounds of the z_1 are $-3.3470,\ 14.9971$, and the lower and upper bounds of the z_2 are $2.1571,\ 15.3812$ respectively. Calculate the two sample means of the reported responses $\bar{z}_1 = \frac{\sum_{i=1}^{n_1} z_i}{n_1}$ and $\bar{z}_2 = \frac{\sum_{i=1}^{n_2} z_i}{n_2}$ for each sample. Then derive the estimate $\hat{\mu}_x^z$ as follows:

$$\hat{\mu}_x^z = \frac{\theta_2 \bar{z}_1 - \theta_1 \bar{z}_2}{\theta_2 - \theta_1} = -1.8154, \quad \theta_1 \neq \theta_2 \tag{4.13}$$

We generate $n_B = 1000$ NPI-Bootstrapping samples $b_1, ...b_{n_B}$ with size n based on the z_i values. Then, calculate the bootstrap sample mean based on the original sample by $\bar{b}_1 = \frac{\sum_{i=1}^{n_1} b_i}{n_1}$ and $\bar{b}_2 = \frac{\sum_{i=1}^{n_2} b_i}{n_2}$ for the two bootstrap samples. Then, derive the estimate $\hat{\mu}_x^B$

based on the NPI-Bootstrap samples as follows:

$$\hat{\mu}_x^B = \frac{\theta_2 \bar{b}_1 - \theta_1 \bar{b}_2}{\theta_2 - \theta_1}, \quad \theta_1 \neq \theta_2 \tag{4.14}$$

Then, we derive the $\epsilon-$reproducibility of the mean as follows:

$$\begin{aligned} \hat{RP}(\epsilon) &= P\biggl(|\hat{\mu}_x^z - \hat{\mu}_x^B| \leq \epsilon\biggr) \Longleftrightarrow P\biggl(\left|\frac{\theta_2 \bar{z}_1 - \theta_1 \bar{z}_2}{\theta_2 - \theta_1} - \frac{\theta_2 \bar{b}_1 - \theta_1 \bar{b}_2}{\theta_2 - \theta_1}\right| \leq \epsilon\biggr) \\ &= \sum_{i=1}^{n_B} \frac{1}{n_B} \mathbf{1}\biggl\{\left|\frac{\theta_2(\bar{z}_1 - \bar{b}_1^i) - \theta_1(\bar{z}_2 - \bar{b}_2^i)}{\theta_2 - \theta_1}\right| \leq \epsilon\biggr\} \end{aligned} \tag{4.15}$$

Repeat this procedure for $n^* = 1000$ times. Figure 4.10 shows the $\epsilon-$reproducibility of the mean of different $\epsilon \in [0, 28]$. It is noticed that $\epsilon-$reproducibility of estimates increases if the ϵ increases, and $\epsilon-$reproducibility based on AM method needs larger $\epsilon = 28$ to obtain higher reproducibility closes to 1.

For more variations, we calculate $\hat{RP}(\epsilon)$ for different ϵ as shown in Figure 4.11 and Table 4.8. As shown in Figure 4.12 and Table 4.9, reproducibility probabilities for estimate decreases if the ψ increases till $\psi = 0.5$, then reproducibility probabilities increase. The highest reproducibiliy of of all characteristics are for $\psi = 0$ and $\psi = 1$. Here all the respondents use the true answers where $\psi = 0$ and then the $\epsilon-$reproducibility of all characteristics are similar to $\epsilon-$reproducibility of all characteristics of $\psi = 1$ where all respondents use scrambling responses. If $\psi = 0.5$, we get the lowest $\epsilon-$reproducibility for responses that divided into two samples. One of them has the responses X, and the other has the responses $X + S_j$.

Example 4.2.5 This example explains reproducibility for estimate based on the additive method (AM) for larger sample size as explained in Section 1.2.2. Assume that a sample of size $n = 500$ is divided into two sub-samples of sizes $n_1 = 279$ and $n_2 = 221$, where $n_1 + n_2 = n$. Let assume true response X_i be simulated from $N(\mu_x = 4, \sigma_x^2 = 3)$. Let $S_1 \sim N(\theta_1 = 3, \gamma_1^2 = 2)$ and $S_2 \sim N(\theta_2 = 5, \gamma_2^2 = 4)$. where $\theta_1 \neq \theta_2$. Suppose the sensitivity level is $\psi = 0.70$.

In this example, it is noted that increasing the sample size leads to increasing the $\epsilon-$reproducibility for the mean AM method as shown in Figures 4.11 and 4.12 and Tables 4.8 and 4.9.

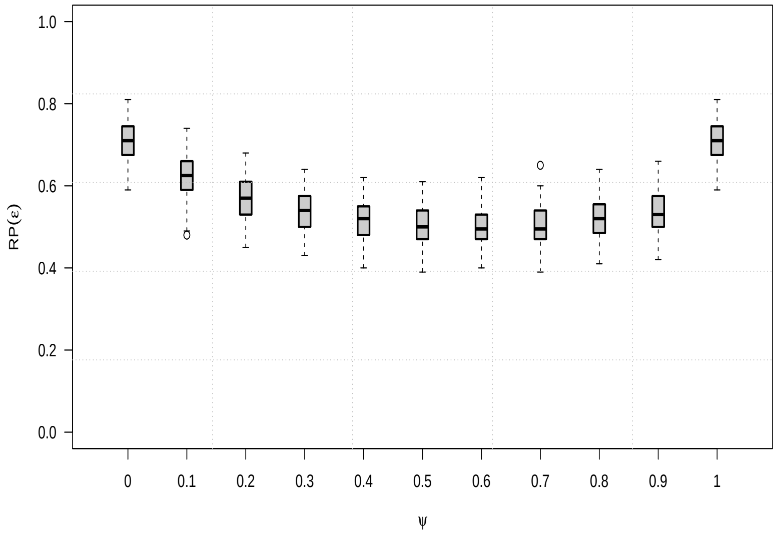


Figure 4.11: $\hat{RP}(\epsilon)$ of the AM method $n = 500,\ n_B = 1000,\ n^* = 100,\ \mu_x = 4,\ \sigma^2_x = 3,$ $\theta_1 = 3,\ \theta_2 = 5,\ \sigma^2_{s_1} = 2,\ \sigma^2_{s_2} = 4,\ \epsilon = 0.5$

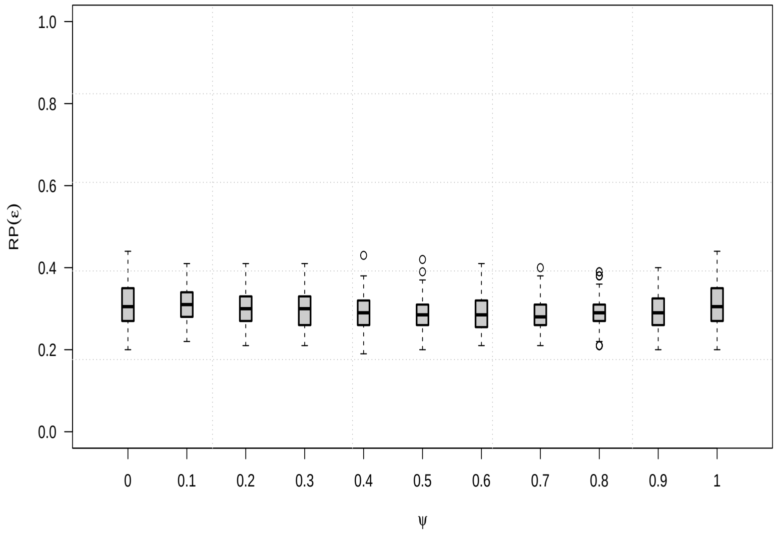


Figure 4.12: $\hat{RP}(\epsilon)$ of AM method $n = 500,\ n_B = 1000,\ n^* = 100,\ \mu_x = 4,\ \sigma^2_x = 20,$ $\theta_1 = 3,\ \theta_2 = 5,\ \sigma^2_{s_1} = 2,\ \sigma^2_{s_2} = 4,\ \epsilon = 1$

Table 4.8 show that the $\hat{RP}(0.5)$ of the mean takes values between 0.50 and 0.71 and the $\hat{RP}(0.5)$ of the median takes value between 0.50 and 0.71. The $\hat{RP}(0.5)$ of $q(0.25)$ takes

Summary	$\psi = 0$	$\psi = 0.1$	$\psi = 0.2$	$\psi = 0.3$	$\psi = 0.4$	$\psi = 0.5$	$\psi = 0.6$	$\psi = 0.7$	$\psi = 0.8$	$\psi = 0.9$	$\psi = 1.0$
$q(0.25)$	0.6775	0.5900	0.5300	0.5000	0.4800	0.4700	0.4700	0.4700	0.4875	0.5000	0.6775
$q(0.75)$	0.7425	0.6600	0.6100	0.5725	0.5500	0.5400	0.5300	0.5400	0.5525	0.5725	0.7425
median	0.7100	0.6250	0.5700	0.5400	0.5200	0.5000	0.4950	0.4950	0.5200	0.5300	0.7100
mean	0.7068	0.6232	0.5675	0.5362	0.5171	0.5057	0.5005	0.5046	0.517	0.5356	0.7068
sd	0.0487	0.0508	0.05428	0.0480	0.0464	0.0508	0.0462	0.0507	0.0495	0.0490	0.0487
IQR	0.0650	0.0700	0.0800	0.0725	0.0700	0.0700	0.0600	0.0700	0.0650	0.0725	0.0650
lowest whisker	0.5800	0.4850	0.4100	0.3913	0.3750	0.3650	0.3800	0.3650	0.3900	0.39125	0.5800
highest whisker	0.8400	0.7650	0.7300	0.6813	0.6550	0.6450	0.6200	0.6450	0.6500	0.6813	0.8400

Table 4.8: $RP(\epsilon)$ of AM method $n = 500$,n$_B = 100,\ n^* = 1000,\ \mu_x = 4,\ \sigma_x^2 = 3,$ $\theta_1 = 3,\ \theta_2 = 5,\ \sigma_{s_1}^2 = 2,\ \sigma_{s_2}^2 = 4,\ \epsilon = 0.5$

Summary	$\psi = 0$	$\psi = 0.1$	$\psi = 0.2$	$\psi = 0.3$	$\psi = 0.4$	$\psi = 0.5$	$\psi = 0.6$	$\psi = 0.7$	$\psi = 0.8$	$\psi = 0.9$	$\psi = 1.0$
$q(0.25)$	0.2700	0.2800	0.2700	0.2600	0.2600	0.2600	0.2575	0.2600	0.2700	0.2600	0.2700
$q(0.75)$	0.3500	0.3400	0.3300	0.3300	0.3200	0.3100	0.3200	0.3100	0.3100	0.3225	0.3500
median	0.3050	0.3100	0.3000	0.3000	0.2900	0.2850	0.2850	0.2800	0.2900	0.2900	0.3050
mean	0.3120	0.3091	0.2998	0.2948	0.2903	0.2882	0.2872	0.2868	0.2898	0.2954	0.3120
sd	0.0527	0.0423	0.04202	0.0422	0.0437	0.0434	0.0436	0.04012	0.0424	0.04300	0.0527
IQR	0.0800	0.0600	0.0600	0.0700	0.0600	0.0500	0.0625	0.0500	0.0400	0.06250	0.0800
lowest whisker	0.1500	0.1900	0.1800	0.1550	0.1700	0.1850	0.16375	0.1850	0.2100	0.1663	0.1500
highest whisker	0.4700	0.4300	0.4200	0.4350	0.4100	0.3850	0.4138	0.3850	0.3700	0.4163	0.4700

Table 4.9: $\hat{RP}(\epsilon)$ of AM method $n = 500$, n$_B = 100,\ n^* = 1000, \mu_x = 4,\ \sigma_x^2 = 20,$ $\theta_1 = 3,\ \theta_2 = 5,\ \sigma_{s_1}^2 = 2,\ \sigma_{s_2}^2 = 4,\ \epsilon = 0.5$

values between 0.47 and 0.67 whereas the $\hat{RP}(0.5)$ of $q(0.75)$ takes values between 0.47 and 0.53. The $\hat{RP}(0.5)$ of the IQR takes values between 0.06 and 0.07. The $\hat{RP}(0.5)$ of the standard deviation sd takes small values between 0.04 and 0.05. The highest whisker takes values between 0.36 and 0.58. and the lowest whisker takes values between 0.62 and 0.84.

It is noticed that if the sensitivity level ψ, the $\epsilon-$reproducibility reduces when $\psi \in [0, 0.5]$ whereas the $\epsilon-$reproducibility decrease of $\psi \in (0.5, 1]$. Increasing the variance of the distribution of the sensitive question leads to decreasing the $\epsilon-$reproducibility as Figure 4.12 and Table 4.9 are shown.

4.3 Reproducibility of estimates using a representative sample

In this section, the $\epsilon-$reproducibility for estimates using the representative sample obtained from a distribution as explained in Section 3.3. This data generated by the RRT methods; EH, MM and GM methods as explained in Section 1.2.2.

Example 4.3.1

This example illustrates $\epsilon-$reproducibility for estimates based on data generated by the EH method. Let's assume that we have a sample with size $n = 3$ and the future sample size is $m = 3$. Assume that the true answer $X_i \sim N(\mu = 4, \sigma^2 = 3)$ be a random quantity as a sensitive characteristic for individual i with an unknown mean μ, and random quantity S_i as a scrambling variable. By giving the randomisation device, we generate random quantity $S_i \sim N(\theta = 1, r^2 = 0.04)$. The random quantities of the original sample is $Z_i \sim N(\mu\theta, r^2\theta^2 + \frac{r^2}{\theta^2}(\sigma^2 + \mu^2)) = \{-0.5865, 4, 8.5865\}$ where $\bar{z}$ is mean of the original sample.

To apply NPI-B method, we calculate the lower and upper bounds as follows: $z_0 = \min(z_i) - d = -5.1731, \quad z_{n+1} = \max(z_i) + d = 13.1731$ and $d = 4.5865$ is the maximal distance between two consecutive values of z_i.

The original sample mean is calculated from $\bar{x} = \frac{\bar{z}}{\theta} = 4$. Here, we have 30 data, so we can construct four intervals between the data set values including the endpoints (z_0, z_{n+1}) with the intervals $I_1 = (1.3842, -0.5865)$, $I_2 = (-0.5865, 4)$, $I_3 = (4, 5.3079)$, and $I_4 = (5.3079, 8.5865)$.

Then, set all possible locations of the future observations as $m = 20$ and the number of the orderings such as $n_o = \binom{n+m}{n} = 20$ to get the orderings O_j. If the order is $(2, 0, 1, 0)$, that means there are two future observations in the interval I_1, no observation in the interval I_2, one observation in I_3, and nothing in the interval I_4 where all orderings have equal probability 1/20.

$\underline{z}^f_{j,l}$	$\epsilon_{j,l}$	$\underline{RP}(\epsilon_{j,l})$	$\overline{z}^f_{j,u}$	$\epsilon_{j,u}$	$\overline{RP}(\epsilon_{j,u})$
1.3842	0.8719	0.35	2.6921	0.0000	0.55
3.5640	0.8719	0.35	4.8719	0.0000	0.55
3.5640	0.8719	0.35	4.8719	0.0000	0.55
2.6921	0.8719	0.35	4.0000	0.0000	0.55
4.4360	0.8719	0.35	5.7438	0.4360	0.80
3.5640	0.8719	0.35	4.8719	0.4360	0.80
2.2562	0.8719	0.35	3.5640	0.4360	0.80
3.5640	1.3079	0.55	4.8719	0.4360	0.80
4.0000	1.3079	0.55	5.3079	0.4360	0.80
3.1281	1.3079	0.55	4.4360	0.8719	0.95
2.6921	1.3079	0.55	4.0000	0.8719	0.95
2.2562	1.7438	0.80	3.5640	0.8719	0.95
1.8202	1.7438	0.80	3.1281	1.3079	1.00
4.8719	1.7438	0.80	6.1798	0.0000	0.55
2.6921	1.7438	0.80	4.0000	0.0000	0.55
4.4360	1.7438	0.80	5.7438	0.0000	0.55
3.5640	2.1798	0.95	4.8719	0.0000	0.55
4.4360	2.1798	0.95	5.7438	0.4360	0.80
1.8202	2.1798	0.95	3.1281	0.4360	0.80
3.5640	2.6158	1.00	4.8719	0.4360	0.80

Table 4.10: The $\underline{z}^f_{j,l}, \overline{z}^f_{j,u}, \epsilon_{j,l}, \epsilon_{j,u}, \underline{RP}(\epsilon_{j,l})$ using the EH method of $n = 3,\ m = 3,\ n_o = 20$, $\mu_x = 4,\ \theta = 1,\ \sigma^2_x = 3,\ r^2 = 0.04,\ \bar{z} = 4$.

We derive the lower and upper mean $\underline{z}^f_{j,l},\ \overline{z}^f_{j,u}$ using Equation (3.7). Then, we calculate $\underline{RP}(\epsilon_{j,l})$ and $\overline{RP}(\epsilon_{j,l})$ where $\epsilon_{j,l}$ is the maximum value between $(\bar{z} - \underline{z}^f_{j,l})$ and $(\overline{z}^f_{j,u} - \bar{z})$, and $\epsilon_{j,u}$ the maximum value between $(\overline{z}^f_{j,l} - \bar{z})$ and $(\bar{z} - \underline{z}^f_{j,u})$ respectively.

Table 4.10 shows the $\underline{z}^f_{j,l}, \overline{z}^f_{j,u}, \epsilon_{j,l}, \epsilon_{j,u}, \underline{RP}(\epsilon_{j,l})$ and $\overline{RP}(\epsilon_{j,l})$. It is noted that the $\epsilon-$reproducibility of the mean increases if the difference between the mean based on the original and future samples increases. The highest values of $\underline{RP}(\epsilon_{j,l})$ are for the orderings (0 1 0 2), (2 0 1 0) and (0 0 0 3) whereas the highest values of $\overline{RP}(\epsilon_{j,u})$ are for the orderings (0 2 0 1), (1 0 1 1) and (2 1 0 0). The ordering of the future observations play important role in the value of the $\epsilon-$reproducibility of the mean.

For more variations, we generate different sample n^* as original samples with size $n = 30$ as shown in Table 4.11. This shows the characteristics of $\epsilon-$reproducibility of the mean using the EH method including the 25^{th}, 50^{th}(mean), and 75^{th} quartiles which are

	$\underline{RP}(1)$			$\overline{RP}(1)$		
n^*	100	500	1000	100	500	1000
$q(0.25)$	0.9230	0.9240	0.9240	0.9830	0.9820	0.9820
$q(0.75)$	0.9330	0.9350	0.9350	0.9880	0.9880	0.9880
median	0.9275	0.9290	0.9290	0.9850	0.9850	0.9850
mean	0.9281	0.9292	0.9290	0.9849	0.9851	0.9850
sd	0.0083	0.0085	0.0084	0.0039	0.0039	0.0038
IQR	0.0100	0.0110	0.0110	0.0050	0.0060	0.0060
lowest whisker	0.9080	0.9075	0.9075	0.9755	0.9730	0.9730
highest whisker	0.9480	0.9515	0.9515	0.9955	0.9970	0.9970

Table 4.11: Estimates of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the EH method of $n = m = 30,\ n_o = 20,$ $\mu_x = 4,\ \theta = 1,\ \sigma_x^2 = 3,\ r^2 = 0.04$

known as the lower quartile $q(0.25)$, median or $q(0.50)$, and upper quartile $q(0.75)$, and then determine the interquartile range IQR.

The results demonstrate that as $n*$ increases, the $\epsilon-$reproducibility increases. There is no difference between the estimates of $n^* = 500$ and $n^* = 1000$ which means more replications lead to more accurate $\epsilon-$reproducibility. The mean and the median takes value 0.92 and 0.98 of the lower and upper $\epsilon-$reproducibilities respectively. The standard deviation is always minimal, indicating that the variance of reproducibility for data points is small and that the distance between data points and the mean is short. The IQR is a measure for determining how far off data points in a set are from the set's mean. It always takes value 0.1001 of $\underline{RP}(\epsilon)$, and takes values between 0.0500 and 0.0600 of $\overline{RP}(\epsilon)$, implying that the smaller the IQR, the more closely the data points are clustered around the mean. The spread data points are quantified using whisker spread.

Table 4.12 shows the average of $\underline{RP}(1)$ and $\overline{RP}(1)$ of different sample sizes and different orderings n_o of $n^* = 100$. It is noted that an increasing number of orderings n_o leads to a decrease in the NPI lower and upper reproducibility probabilities. Increasing of the sample size leads to higher NPI lower and upper reproducibility probabilities which means we obtain more accurate information about the $\epsilon-$reproducibility.

Example 4.3.2 This example illustrates $\epsilon-$reproducibility of the mean of MM method [62] of a sample with size $n = 3$. Suppose that the true answer $X_i \sim N(\mu_x = 4, \sigma_x^2 = 3)$ be a random quantity as a sensitive characteristic for individual i with an unknown mean μ_x,

n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$	n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$
5	100	0.4684	0.9049	100	100	0.9998	1.0000
	500	0.4550	0.9061		500	0.9994	0.9999
	1000	0.4539	0.9047		1000	0.9994	0.9998
n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$	n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$
30	100	0.9301	0.9854	1000	100	1.0000	1.0000
	500	0.9278	0.9849		500	1.0000	1.0000
	1000	0.9289	0.9848		1000	1.0000	1.0000

Table 4.12: The average of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the EH method of different n, n_o of $n^* = 100\ \mu_x = 4,\ \theta = 1,\ \sigma_x^2 = 3,\ r^2 = 0.04$

and random quantity S_i as a scrambling variable. By giving the randomisation device, we generate a random quantity $S_i \sim N(\theta = 1, \gamma^2 = 0.04)$ where the sensitivity level of the sensitive question is $\psi = 0.7$.

We generate the original generating sample $Z_i \sim N(\mu_x, \sigma_x^2+\psi\gamma^2(\sigma_x^2+\mu_x^2)) = N(4, 3.532)$, where $i = 1, ..., n$ such that $z_1 = 2.7324,\ z_2 = 4$, and $z_3 = 5.2676$. where the mean of the original sample is $\bar{z} = 4$. To apply the NPI-B method, we determine the lower and upper bounds $z_0 = 1.4648$, $z_{n+1} = 6.5352$ where $d = 1.2676$. We generate the orderings of the future observations $\binom{6}{3} = 20$. Then calculate the lower and upper mean $\underline{z}_l^f$, $\overline{z}_u^f$, and determine the values of $\epsilon_{j,l}$, $\epsilon_{j,u}$ to derive $\underline{RP}(\epsilon)$ and $\overline{RP}(\epsilon)$.

It is noted that a larger distance between the original sample means and lower means $\underline{z}_l^f$ and upper means $\overline{z}_u^f$ leads to larger reproducibility of the mean. Therefore, largest value of $\epsilon_{j,l} = 2.5352$ leads to largest value of $\underline{RP}(2.5352) = 1$. Similarly, the largest value of $\epsilon_{j,u} = 1.2676$ leads to the largest value of $\overline{RP}(1.2676) = 1$ as shown in Table 4.13.

Table 4.14 summarised the characteristics of the MM method including the lower quartile $q(0.25)$, median M or $q(0.50)$, and upper quartile $q(0.75)$ and the interquartile range IQR. The results show that as n^* increases then the lower and upper reproducibility decrease. The means are close to the median. The standard deviation is always minimal, indicating that the variance of $\underline{RP}(\epsilon)$ data points is small and that the distance between data points and the mean is small.

For different replications $n^* = 100,\ 500$ and 1000. The median takes value 0.95 and

$\underline{z}^f_{j,l}$	$\epsilon_{j,l}$	$\underline{RP}(\epsilon_{j,l})$	$\overline{z}^f_{j,u}$	$\epsilon_{j,u}$	$\overline{RP}(\epsilon_{j,u})$
1.4648	0.8451	0.35	2.7324	0.0000	0.55
3.5775	0.8451	0.35	4.8451	0.0000	0.55
3.5775	0.8451	0.35	4.8451	0.0000	0.55
2.7324	0.8451	0.35	4.0000	0.0000	0.55
4.4225	0.8451	0.35	5.6901	0.4225	0.70
3.5775	0.8451	0.35	4.8451	0.4225	0.70
2.3099	0.8451	0.35	3.5775	0.4225	0.70
3.5775	1.2676	0.55	4.8451	0.4225	0.80
4.0000	1.2676	0.55	5.2676	0.4225	0.80
3.1549	1.2676	0.55	4.4225	0.8451	0.90
2.7324	1.2676	0.55	4.0000	0.8451	0.90
2.3099	1.6901	0.80	3.5775	0.8451	0.95
1.8873	1.6901	0.80	3.1549	1.2676	1.00
4.8451	1.6901	0.80	6.1127	0.0000	0.55
2.7324	1.6901	0.80	4.0000	0.0000	0.55
4.4225	1.6901	0.80	5.6901	0.0000	0.55
3.5775	2.1127	0.95	4.8451	0.0000	0.55
4.4225	2.1127	0.95	5.6901	0.4225	0.70
1.8873	2.1127	0.95	3.1549	0.4225	0.70
3.5775	2.5352	1.00	4.8451	0.4225	0.70

Table 4.13: The $\underline{z}^f_{j,l}$, $\overline{z}^f_{j,u}$, $\epsilon_{j,l}$, $\epsilon_{j,u}$, $\underline{RP}(\epsilon_{j,l})$ and $\overline{RP}(\epsilon_{j,u})$ of the MM method of $n = 3$, $\mu_x = 4$, $n_o = 20$, $\theta = 1$, $\sigma^2_x = 3$, $\gamma^2 = 0.04$, $\psi = 0.70$, $\bar{z} = 4$.

1 of the lower and upper $\epsilon-$reproducibilities respectively. The $\underline{RP}(1)$ of the mean takes values between 0.93 and 0.94 whereas $\overline{RP}(1)$ takes value 0.98. The standard deviation is always minimal, indicating that the variance of reproducibility for data points is small and that the distance between data points and the mean is small. The IQR is a measure for determining how far off data points in a set are from the set's mean. It always takes value 0.1000 of $\underline{RP}(1)$, and takes values between 0 of $\overline{RP}(1)$, implying that the smaller the IQR, the more closely the data points are clustered around the mean. The spread data points are quantified using whisker spread.

Increasing sample size leads to increasing the average of the related lower and upper $\epsilon-$reproducibility as shown in Table 4.15. The highest value of lower and upper reproducibility of $n = 1000$ where the lower and upper $\epsilon-$reproducibility are equal. For different orderings n_o, the highest value of lower and upper reproducibility of $n_o = 100$.

	$\underline{RP}(1)$			$\overline{RP}(1)$		
n^*	100	500	1000	100	500	1000
$q(0.25)$	0.9000	0.9000	0.9000	1	1	1
$q(0.75)$	1	1	1	1	1	1
Median	0.9500	0.9500	0.9500	1	1	1
Mean	0.9395	0.9402	0.9360	0.9885	0.9879	0.9871
sd	0.0547	0.0508	0.05512	0.0223	0.0237	0.0246
IQR	0.1000	0.1000	0.1000	0	0	0
lowest whisker	0.7500	0.7500	0.7500	1	1	1
highest whisker	1	1	1	1	1	1

Table 4.14: Estimates of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the MM method of $n = 30,\ n_o = 20,\ \mu_x = 4,\ \theta = 1,\ \sigma_x^2 = 3,\ \gamma^2 = 0.04, \psi = 0.7$

n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$	n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$
	100	0.4684	0.9049		100	1.0000	1.0000
5	500	0.4550	0.9061	100	500	0.9996	0.9999
	1000	0.4539	0.9047		1000	0.9995	0.9999
	100	0.9402	0.9879		100	1.0000	1.0000
30	500	0.9374	0.9874	1000	500	1.0000	1.0000
	1000	0.9384	0.9875		1000	1.0000	1.0000

Table 4.15: Average of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the MM method of different $n,\ n_o$ of $n^* = 100\ of\ \mu_x = 4,\ \theta = 1,\ \sigma_x^2 = 3,\ \gamma^2 = 0.04, \psi = 0.70$

For larger sample size and ordering numbers, the lower and upper $\epsilon-$reproducibility of the mean of the MM method is slightly larger than the lower and upper $\epsilon-$reproducibility of the mean of the EH method at $\psi = 1$.

Example 4.3.3 In this example, we derive the lower and upper reproducibility for estimates based on the GM method [61] using a representative sample with a size of $n = 3$. Assume the probability of the sensitive question is $\gamma = 0.70$. We simulate the random quantity of the responses to the sensitive question $X_i \sim N(\mu_x = 4, \sigma_x^2 = 3)$ and the responses to the unrelated question $Y_i \sim N(\mu_y = 1, \sigma_y^2 = 0.04)$.

We generate random quantities of the original sample Z_i where, $Z_i \sim N(\gamma\mu_y + (1 - \gamma)\mu_x, \sigma_y^2 + \gamma(\sigma_x^2 - \sigma_y^2) + \gamma(1-\gamma)(\mu_x - \mu_y)^2) = N(3.1, 8.1673)$. The first response is $z_1 = 1.1724$, the second response is $z_2 = 3.1$, and the third one is $z_3 = 5.0276$, where the mean of the

$\underline{z}^f_{j,l}$	$\epsilon_{j,L}$	$\underline{RP}(\epsilon_{j,l})$	$\overline{z}^f_{j,u}$	$\epsilon_{j,u}$	$\overline{RP}(\epsilon_{j,u})$
-0.7552	1.2851	0.35	1.1724	0.0000	0.55
2.4575	1.2851	0.35	4.3851	0.0000	0.55
2.4575	1.2851	0.35	4.3851	0.0000	0.55
1.1724	1.2851	0.35	3.1000	0.0000	0.55
3.7425	1.2851	0.35	5.6701	0.6425	0.70
2.4575	1.2851	0.35	4.3851	0.6425	0.70
0.5299	1.2851	0.35	2.4575	0.6425	0.70
2.4575	1.9276	0.55	4.3851	0.6425	0.80
3.1000	1.9276	0.55	5.0276	0.6425	0.80
1.8149	1.9276	0.55	3.7425	1.2851	0.95
1.1724	1.9276	0.55	3.1000	1.2851	0.95
0.5299	2.5701	0.80	2.4575	1.2851	0.95
-0.1127	2.5701	0.80	1.8149	1.9276	1.00
4.3851	2.5701	0.80	6.3127	0.0000	0.55
1.1724	2.5701	0.80	3.1000	0.0000	0.55
3.7425	2.5701	0.80	5.6701	0.0000	0.55
2.4575	3.2127	0.95	4.3851	0.0000	0.55
3.7425	3.2127	0.95	5.6701	0.6425	0.70
-0.1127	3.2127	0.95	1.8149	0.6425	0.70
2.4575	3.8552	1.00	4.3851	0.6425	0.70

Table 4.16: The $\underline{z}^f_{j,l}$, $\overline{z}^f_{j,u}$, $\epsilon_{j,l}$, $\epsilon_{j,u}$,$\underline{RP}(\epsilon_{j,l})$ and $\underline{RP}(\epsilon_{j,u})$ of the GM method of $n = 3$, $n_o = 20$, $\mu_x = 4$, $\mu_y = 1$, $\sigma^2_x = 3$, $\gamma = 0.7$, $\sigma^2_y = 0.04$, $\bar{z} = 3.1$

original sample is $\bar{z} = 3.1$. The lower and upper bounds are $z_0 = -0.7552$ and $z_{n+1} = 6.9552$, where $d = 1.9276$ is the maximal distance between two consecutive of z_i values.

Then, find all possible orderings of the future observations to calculate the lower and upper future averages $\underline{z}^f_l$ and $\overline{z}^f_u$, then, and calculate the maximum values of $\epsilon_{j,l}$ and $\epsilon_{j,u}$, respectively, to derive the lower and upper $\epsilon-$reproducibility of the difference between the average of the original sample and the future sample as Table 4.16 is shown. It is noted that a larger distance between the original sample means and lower means $\underline{z}^f_l$ and upper means $\overline{z}^f_u$ leads to larger reproducibility of the mean. Therefore, largest value of $\epsilon_{j,l} = 3.8552$ leads to largest value of $\underline{RP}(3.8552) = 1$. Similarly, the largest value of $\epsilon_{j,u} = 1.9276$ leads to the largest value of $\overline{RP}(1.9276) = 1$ and the lowest value of $\epsilon_{j,u} = 0$ leads to the largest value of $\overline{RP}(0) = 0.55$ as Table 4.16 is shown.

Table 4.17 shows the lower quartile $q(0.25)$, median and $q(0.25)$, the upper quartile

	$\underline{RP}(1)$			$\overline{RP}(1)$		
n^*	100	500	1000	100	500	1000
$q(0.25)$	0.7000	0.7000	0.7000	0.8500	0.8500	0.900
$q(0.75)$	0.8000	0.8000	0.8000	0.9500	0.9500	0.9500
Median	0.7500	0.7500	0.7500	0.9000	0.9000	0.9000
Mean	0.7415	0.7506	0.7480	0.9155	0.9181	0.9152
sd	0.0935	0.0871	0.0932	0.0610	0.0595	0.0619
IQR	0.1000	0.1000	0.1000	0.1000	0.1000	0.0500
lowest whisker	0.5500	0.5500	0.5500	0.7000	0.7000	0.8250
highest whisker	0.9500	0.9500	0.9500	1	1	1

Table 4.17: Estimates of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the GM method of $n = m = 30$, $n_o = 20,\ \mu_x = 4,\ \mu_y = 1,\ \gamma = 0.7,\ \sigma_x^2 = 3,\ \sigma_y^2 = 0.04$

n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$	n	n_o	Average of $\underline{RP}(1)$	Average of $\overline{RP}(1)$
5	100	0.1729	0.7920	100	100	0.9807	0.9948
	500	0.1682	0.7865		500	0.9799	0.9936
	1000	0.1684	0.7853		1000	0.9795	0.9935
30	100	0.7506	0.9181	1000	100	1.0000	1.0000
	500	0.7504	0.9176		500	1.0000	1.0000
	1000	0.7512	0.9186		1000	1.0000	1.0000

Table 4.18: Average of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the GM method of $n^* = 100$, $\mu_x = 4, \mu_y = 1, \gamma = 0.7,\ \sigma_x^2 = 3,\ \sigma_y^2 = 0.04$

$q(0.75)$, the interquartile range IQR, the mean, the lowest and highest whiskers of the lower and upper reproducibility probabilities for different sample size $n^* = 100,\ 500,\ 1000$. The median does not change with different n^*. The mean takes a value between 0.74 and 0.75. The lowest values are for standard deviation and IQR of the upper reproducibilities.

Table 4.17shows that the increasing of the replication numbers n^* does not affect considerably on the lower and upper $\epsilon-$reproducibily of the mean. That means the increasing in n^* leads to a slight increase in $\underline{RP}(1)$ and $\overline{RP}(1)$.

For different sample sizes and orderings numbers, Table 4.18 shows that increasing sample size leads to increasing the average of the lower and upper $\epsilon-$reproducibility as shown. The lower and upper $\epsilon-$reproducibility of the mean of the GM method is the smallest $\epsilon-$reproducibility.

n_o	Average $\underline{RP}(1)$	$\underline{CI}(0.95)$	Average $\overline{RP}(1)$	$\overline{CI}(0.95)$
1000	0.9273	(0.9112,0.9434)	0.9845	(0.9768,0.9922)
2000	0.9289	(0.9176,0.9402)	0.9849	(0.9796,0.9902)
5000	0.9287	(0.9216,0.9358)	0.9849	(0.9815,0.9883)
10000	0.9290	(0.9240,0.9340)	0.9849	(0.9825,0.9873)
20000	0.9287	(0.9251, 0.9323)	0.9848	(0.9831,0.9865)
50000	0.9290	(0.9267, 0.9313)	0.9848	(0.9837,0.9859)
100000	0.9289	(0.9273, 0.9305)	0.9848	(0.9840,0.9856)

Table 4.19: The lower and upper of CI(95%) of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the EH method of $n = m = 30,\ n^* = 100,\ \mu_x = 2,\ \theta = 0.3,\ \sigma_x^2 = 3,\ \gamma^2 = 0.04$

n_o	Average $\underline{RP}(1)$	$\underline{CI}(0.95)$	Average $\overline{RP}(1)$	$\overline{CI}(0.95)$
1000	0.9370	(0.9219, 0.9521)	0.9871	(0.9801,0.9941)
2000	0.9384	(0.9279, 0.9489)	0.9875	(0.9826,0.9924)
5000	0.9380	(0.9313, 0.9447)	0.9875	(0.9844,0.9906)
10000	0.9382	(0.9335, 0.9429)	0.9875	(0.9853,0.9897)
20000	0.9381	(0.9348, 0.9414)	0.9875	(0.9860,0.9890)
50000	0.9384	(0.9363, 0.9405)	0.9874	(0.9864,0.9884)
100000	0.9384	(0.9369, 0.9399)	0.9874	(0.9867,0.9881)

Table 4.20: The lower and upper of CI(95%) of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the MM method of $n = m = 30,\ n^* = 100,\ \mu_x = 2,\ \theta = 0.3,\ \sigma_x^2 = 3,\ \gamma^2 = 0.04,\ \psi = 0.70$

As a result, the sample size and choosing parameters of the RRT methods have a basic role to obtain high the lower and upper $\epsilon-$reproducibility of RRT method.

Now, we computed the exact lower and upper $\epsilon-$reproducibility probabilities for a sample with a size of $n = m$, considering large orderings n_o using the SOM methodology to compute the lower and upper $\epsilon-$reproducibility probabilities in order to assess the precision of the SOM method of the computation of the lower and upper $\epsilon-$reproducibility of RRT methods. The Normal distribution is assumed to be the underlying distributions. Then, we generate n_o orderings equal to $1000, 2000, 5000, 10000, 20000, 50000, 100000$ using the SOM approach, and the 95% confidence interval was calculated for both the lower and upper bounds in each replication as shown on Tables 4.19, 4.20 and 4.21

To calculate the lower confidence interval of exact $\epsilon-$reproducibility for estimates where the number of sampled orderings is larger than or equal 1000, the interval is calculated

n_o	Average $\underline{RP}(1)$	$\underline{CI}(0.95)$	Average $\overline{RP}(1)$	$\overline{CI}(0.95)$
1000	0.7512	(0.7244, 0.7780)	0.9186	(0.9017, 0.9355)
2000	0.7511	(0.7322 , 0.7700)	0.9192	(0.9073, 0.9311)
5000	0.7501	(0.7311, 0.7691)	0.9186	(0.9066, 0.9306)
10000	0.7497	(0.7307, 0.7687)	0.9184	(0.9064, 0.9304)
20000	0.7496	(0.7436, 0.7556)	0.9184	(0.9146, 0.9222)
50000	0.7494	(0.7456, 0.7532)	0.9183	(0.9159, 0.9207)
100000	0.7497	(0.7470, 0.7524)	0.9184	(0.9167, 0.9201)

Table 4.21: The lower and upper of CI(95%) of $\underline{RP}(1)$ and $\overline{RP}(1)$ using the GM method of $n = m = 30,\ n^* = 100,\ \mu_x = 4,\ \mu_y = 1,\ \gamma = 0.7,\ \sigma_x^2 = 3,\ \sigma_y^2 = 0.04$

using the normal approximation for the ordering number n_o. Therefore, the confidence intervals of the $\underline{RP}(\epsilon)$ and $\overline{RP}(\epsilon)$ are

$$\underline{RP}(\epsilon) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\underline{RP}(\epsilon)\left(1 - \underline{RP}(\epsilon)\right)}{n_o}}, \qquad \overline{RP}(\epsilon) \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\overline{RP}(\epsilon)\left(1 - \overline{RP}(\epsilon)\right)}{n_o}} \tag{4.16}$$

where $z_{\frac{\alpha}{2}}$ is $1 - \frac{\alpha}{2}$ quantile of the standard Normal distribution.

Tables 4.19, 4.20 and 4.21 show that the $\underline{RP}(\epsilon)$ and $\overline{RP}(\epsilon)$ of the mean of GM method have the smallest $\epsilon-$reproducibility whereas the $\underline{RP}(\epsilon)$ and $\overline{RP}(\epsilon)$ of the mean of the MM method have the largest $\epsilon-$reproducibility. Increasing n_o leads to a slight decreases in the $\underline{RP}(\epsilon)$ and $\overline{RP}(\epsilon)$ of the mean of RRT methods and the confidence intervals $\underline{CI}(95\%)$ and $\overline{CI}(95\%)$.

In general, it is noted that increasing the number of orderings of future observations leads to an increase in the approximate lower and upper reproducibility for estimates based on RRT and decreases the range of lower and upper confidence intervals for these lower and upper reproducibility probabilities.

4.4 Comparison of RRT methods

We compare RRT methods for real-valued quantities based on three properties; variance, privacy degree, and $\epsilon-$reproducibility of estimates. To compare the $\epsilon-$reproducibility of estimates based on MM, EH, and GM data using simulation with NPI-B and the

	$n = m =$	100	300	500	Δ_{GM}	$\text{Var}(Z_i)$
$\gamma = 0.7$	$RP(\epsilon)$	0.3330	0.4730	0.5690	1.2234	4.5376
	$\underline{RP}(\epsilon)$	0.1330	0.3680	0.5020		
	$\overline{RP}(\epsilon)$	0.4170	0.5260	0.6170		
$\gamma = 0.53$	$\hat{RP}(\epsilon)$	0.2460	0.3680	0.4450	1.9167	7.3573
	$\underline{RP}(\epsilon)$	0.0680	0.2770	0.3840		
	$\overline{RP}(\epsilon)$	0.3740	0.4570	0.5280		

Table 4.22: Reproducibility of the mean of GM of $n_B = 1000$, $n_o = 1000$, $\mu_x = 4$, $\mu_y = 4$, $\sigma_x^2 = 2.5$, $\sigma_y^2 = 1.5780$, $\epsilon = 0.1$

representative sample, we set parameters values for each method to achieve the same privacy degree, then assess the variance and reproducibility of each method.

We first fix some parameters values of RRT methods such as the sample size $n = 100,\ 300,\ 500$, the number of ordering $n_o = 1000$, the number of NPI-B samples $n_B = 1000$ and $\epsilon = 0.1,\ 0.7$. Then we set the other parameter values to obtain the same privacy degree. For example; the parameter values of the GM methods are $\mu_x = 4$, $\mu_y = 4$, $\sigma_x^2 = 2.5$, $\sigma_y^2 = 1.5780$ and $\gamma = 0.70$. The parameter values of the MM method are $\mu_x = 2$, $\theta = 1$, $\sigma_x^2 = 4.8$, $\gamma^2 = 0.2958$, and $\psi = 0.70$. The parameter values of the EH methods are $\mu_x = 4$, $\theta = 2$, $\sigma_x^2 = 12.543$, $r^2 = 0.2958$ to obtain the same privacy degree 1.2234. Then, we change γ, ψ and ϵ to investigate the changes in the RRT method in terms of privacy degree, the variance and the $\epsilon-$reproducibility of the mean.

Tables 4.22, 4.23 and 4.24 show that the comparison between GM, MM and EH methods using the NPI-B and the representative sample. The results show that the lower and upper $\epsilon-$reproducibility of the mean of RRT methods increases if the variance decreases (the efficiency of the method increases) while the privacy degree decreases.

Tables 4.22, 4.23 and 4.24 show that the $\epsilon-$reproducibility of the mean based on the GM, MM and EH method increases if γ increases or the sample size increases. The $\epsilon-$reproducibility for an estimate using NPI-B gets values within the range of the lower and upper reproducibility of RRT using the representative sample except for the cases in which the difference between the mean $\hat{\mu}_x^z$ of original samples of the reproducibility using the NPI-B method are large than the mean $\hat{\mu}_x^z$ of original samples of the reproducibility using the representative sample more than 0.2.

	$n = m =$	100	300	500	Δ_{MM}	Var(Z_i)
MM($\psi = 0.70$)	$RP(\epsilon)$	0.2850	0.4220	0.5010	0.7809	6.6221
	$\underline{RP}(\epsilon)$	0.0830	0.2890	0.4170		
	$\overline{RP}(\epsilon)$	0.3830	0.4680	0.5410		
MM($\psi = 0.53$)	$RP(\epsilon)$	0.2850	0.4220	0.5010	1.2234	6.1796
	$\underline{RP}(\epsilon)$	0.0900	0.3100	0.4260		
	$\overline{RP}(\epsilon)$	0.3900	0.4780	0.5520		

Table 4.23: Reproducibility of the mean of MM of $n_B = 1000$, $n_o = 1000$, $\mu_x = 2$, $\theta = 1$, $\sigma_x^2 = 4.8$, $\gamma^2 = 0.2958$, $\epsilon = 0.1$

EH	$n = m =$	100	300	500	Δ_{EH}	Var(Z_i)
	$RP(\epsilon)$	0.1160	0.1900	0.2340	1.2234	13.7664
	$\underline{RP}(\epsilon)$	0.0160	0.1770	0.2750		
	$\overline{RP}(\epsilon)$	0.3270	0.3580	0.4140		

Table 4.24: Reproducibility of the mean of EH of $n_B = 1000$, $n_o = 1000$, $\mu = 2$, $\theta = 2$, $\sigma_x^2 = 12.543$, $r^2 = 0.2958$, $\epsilon = 0.1$

Increasing the sample size leads to higher $\epsilon-$reproducibility and obtains higher values of lower and upper $\epsilon-$reproducibility using the representative sample and $\epsilon-$reproducibility using NPI-B method. Tables 4.23 and 4.24 show that $\epsilon-$reproducibility of the mean based on MM is higher than the RP for estimates based on EH. Increasing ϵ leads to an increase of $\epsilon-$reproducibility as shown in Tables 4.25, 4.26 and 4.27.

Based on the comparisons of the quantitative RRT methods, it is observed that at the same level of privacy protection. The privacy degree, the variance and the $\epsilon-$reproducibility of the mean of the EH method are equivalent to the privacy degree, the variance of Z_i and the $\epsilon-$reproducibility of the mean of the MM method of the sensitivity level $\psi = 1$. The GM method has less variability of the reported responses than the EH and the MM method at the same privacy degree.

To conclude, the $\epsilon-$reproducibility of estimates is affected by the variance of the original sample (the variability in the reported responses). If the variance increases (the variability of the reported responses are large), then the $\epsilon-$reproducibility of estimates decreases. Higher $\epsilon-$reproducibility of estimates leads to lower privacy degree of the RRT methods.

	$n=m=$	100	300	500	Δ_{GM}	$\text{Var}(Z_i)$
$\gamma = 0.7$	$\hat{RP}(\epsilon)$	0.9890	1.0000	1.0000	1.2234	4.5376
	$\underline{RP}(\epsilon)$	0.9750	1.0000	1.0000		
	$\overline{RP}(\epsilon)$	0.9900	1.0000	1.0000		
$\gamma = 0.53$	$RP(\epsilon)$	0.9270	1.0000	1.0000	1.9167	7.35728
	$\underline{RP}(\epsilon)$	0.9150	0.9990	1.0000		
	$\overline{RP}(\epsilon)$	0.9670	1.0000	1.0000		

Table 4.25: Reproducibility of the mean of GM of $n_B = 1000$, $n_o = 1000$, $\mu_x = 4$, $\mu_y = 4$, $\sigma_x^2 = 2.5$, $\sigma_y^2 = 1.5780$, $\epsilon = 0.7$

	$n=m=$	100	300	500	Δ_{MM}	$\text{Var}(Z_i)$
MM($\psi = 0.70$)	$\hat{RP}(\epsilon)$	0.9850	1.0000	1.0000	0.7809	6.6221
	$\underline{RP}(\epsilon)$	0.9270	1.0000	1.0000		
	$\overline{RP}(\epsilon)$	0.9740	1.0000	1.0000		
MM($\psi = 0.53$)	$\hat{RP}(\epsilon)$	0.9850	1.0000	1.0000	1.2234	6.1796
	$\underline{RP}(\epsilon)$	0.9370	1.0000	1.0000		
	$\overline{RP}(\epsilon)$	0.9780	1.000	1.0000		

Table 4.26: Reproducibility of the mean of MM of $n_B = 1000$, $n_o = 1000$, $\mu_x = 2$, $\theta = 1$, $\sigma_x^2 = 4.8$, $\gamma^2 = 0.2958$, $\epsilon = 0.7$

EH	$n=m=$	100	300	500	Δ_{EH}	$\text{Var}(Z_i)$
	$\hat{RP}(\epsilon)$	0.6990	0.8970	0.9640	1.2234	13.7664
	$\underline{RP}(\epsilon)$	0.7750	0.9740	0.9940		
	$\overline{RP}(\epsilon)$	0.8910	0.9890	0.9970		

Table 4.27: Reproducibility of the mean of EH of $n_B = 1000$, $n_o = 1000$, $\mu = 2$, $\theta = 2$, $\sigma_x^2 = 12.543$, $r^2 = 0.2958$, $\epsilon = 0.7$

4.5 Concluding remarks

This chapter studies $\epsilon-$reproducibility of estimates as introduced in Chapter 3 based on quantitative RRT methods in two ways; the first method uses NPI-B method and the other method uses the representative sample.

This first method investigates the $\epsilon-$reproducibility of estimates based on the randomised response method by using the simulation. The $\epsilon-$reproducibility of estimates has different behaviour depending on the design of RRT methods.

Using NPI-Bootstrap method is an excellent procedure to generate all possible future observations of the original sample while SOM method is a helpful technique to generate all possible orderings of future observations. Therefore, for a large sample size, if it cannot consider all the orderings of the future observations, we use sampling of ordering method (SOM) to obtain a large number of orderings to derive approximation of the lower and upper $\epsilon-$reproducibility. Using a larger sample size n leads to a decrease in the difference between the lower and upper $\epsilon-$reproducibility and gives accurate $\epsilon-$reproducibility. A lower variance of the reported responses leads to higher $\epsilon-$reproducibility with the same privacy degree.

It is noted that $\epsilon-$reproducibility of an estimate of the GM method has less variability of the reported responses than the MM and EH methods. There is a strong relationship between this variability and higher $\epsilon-$reproducibility of estimates of RRT method. Less variability leads to high $\epsilon-$reproducibility of an estimate. Increasing ϵ and the sample size n leads to higher $\epsilon-$reproducibility of an estimate.

For further research, this work can be applied to different RRT methods that have different procedures or multiple samples. In addition, $\epsilon-$reproducibility of estimates can be improved to investigate a unified measure to connect the variability of the reported responses, respondents' privacy and $\epsilon-$reproducibility of estimates.

Chapter 5

Conclusions

In this thesis, we presented the reproducibility probability for hypothesis test scenarios with data collected using RRT methods. We further proposed a novel method to study reproducibility of estimates, and we applied this to compare different RRT methods.

In Chapter 2, the reproducibility of the statistical hypothesis test was presented for data derived from two types of randomised response methods: the Greenberg model and the forced methods. This reproducibility of the statistical hypothesis tests is applied for one-sided and two-sided hypothesis tests of the proportion of the respondents who response 'yes'. Besides, a new measurement of reproducibility is proposed to compare the RRT methods.

This method does not work well with larger samples n. It will be interesting, for further research, to study the reproducibility of statistical tests based on RRT methods considering larger sample sizes. We could also use future sample sizes that differ from the data sample. The results show that the forced method has less variability of the reported responses and higher reproducibility with the same privacy degree.

In Chapter 3, we discussed the $\epsilon-$reproducibility of estimates of real data is generating from the standard normal distribution. This method is applied using the NPI-Bootstrap. We use NPI-Bootstrap to generate new samples of the future observation and then to estimate the population characteristics. We obtain $\epsilon-$reproducibility by finding the probability of an estimate valued future observations is close to the actual estimates.

The other procedure for this method we proposed the representative sample as a new approach to generate the original sample, we consider the ordering of the future sample, and we obtain the exact estimates which are close to the actual estimates. That can be applied easily for the mean, the median, the variance, the quartiles, and IQR. However, we can implement all the lower and upper $\epsilon-$reproducibility of estimates of these characteristics except the lower and upper variance in the case of the $\epsilon-$reproducibility for estimation by using the representative sample. Therefore, it may be well to search for this in the future because the upper variance is a quadratic constraint optimisation problem to which solutions are normally not available in closed-form.

Chapter 4 investigates the reproducibility of point estimates of population characteristics based on data collected by RRT methods such as the Greenberg method, the multiplicative method, and the additive optional method. The results show that the $\epsilon-$reproducibility of estimates of the Greenberg method is higher than the other RRT methods. In general, We find that less variability in the reported responses of RRT methods leads to higher reproducibility with the same privacy degree. In this chapter, we choose the method that is simple to apply, not because it is the most essential method in the practical way to assess the reproducibility of statistical inference based on RRT methods. It will be crucial for the upcoming research to examine a wide range of RRT techniques, including the additive models and or the combination of additive and multiplicative models.

Finally, applying reproducibility based on RRT will be a great idea if we investigate the following ideas. It is useful to investigate a unified measurement over the fixed privacy level and reproducibility and link this work with Gupta et.al work [63] and compare or combine it with the unified measure of privacy level and efficiency. In addition, it is essential to investigate reproducibility using different statistical inference techniques that can be appropriate for reproducibility probability. Furthermore, it is a good idea to apply the reproducibility method for a range of further statistical inferences based on RRT including multiple-sample scenarios.

Bibliography

[1] Alabdulhadi, M., Coolen-Maturi, T. and Coolen, F.P.A. (2021). Nonparametric predictive inference for comparison of two diagnostic tests. *Communications in Statistics - Theory and Methods*, **19**, 4470-4486.

[2] Abernathy, J. R., Greenberg, B. G. and Horvitz, D. G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, **7**, 19–29

[3] Aboalkhair, A.M. (2012). *Nonparametric predictive inference for system reliability.* PhD Thesis. Durham University. Available at: `http://npi-statistics.com`.

[4] Adebola, F. B., Adediran, A. A. and Ewemooje, O. S. (2017). Hybrid tripartite randomized response technique. *Communications in Statistics–Theory and Methods*, **46**, 11756–11763.

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

[6] Agresti, A. and Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, **54**, 280–288.

[7] Alqifari, H.N. (2017). *Nonparametric Predictive Inference for Future Order Statistics.* PhD Thesis. Durham University. Available at : `http://npi-statistics.com`.

[8] Alqifari, H.N. and Coolen, F.P.A. (2019). Robustness of nonparametric predictive inference for future order statistics. *Journal of Statistical Theory and Practice*, **4**, 12.

[9] Anderson, H. (1977). Efficiency versus protection in a general randomized response model. *Scandinavian Journal of Statistics*, **4**, 11–19.

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

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

[12] Atmanspacher, H. and Maasen, S. (eds.) (2016). *Reproducibility: Principles, Problems, Practices and Prospects*. New Jersey:Wiley.

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

[14] Augustin, T. Coolen, F.P.A. De Cooman, G. and Troffaes, M.C.(eds.).(2014). *Introduction to Imprecise Probabilities.* Chichester: John Wiley & Sons.

[15] Baker, R.M., Coolen-Maturi, T. and Coolen, F.P.A. (2017). Nonparametric predictive inference for stock returns, *Journal of Applied Statistics*, **44**(8), 1333– 1349.

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

[17] Barton, B. and Peat, J. (2014). *Medical Statistics: A Guide to SPSS, Data Analysis and Critical Appraisal.* Oxfors: John Wiley & Sons.

[18] Bin Himd, S. (2014). *Nonparametric predictive methods for bootstrap and test reproducibility.* PhD Thesis. Durham University. Available at: `http://npi-statistics.com`.

[19] Chaudhuri, A. (2016). Randomized response and indirect questioning techniques in surveys. New York: CRC.

[20] Chen, J., Coolen, F.P.A. and Coolen-Maturi, T. (2019) On nonparametric predictive inference for asset and European option trading in the binomial tree model. *Journal of the Operational Research Society*, **70**, 1678-1691.

[21] Chow, S., Shao, J. and Wang H. (2008). *Sample Size Calculations in Clinical Research.* Second Edition. New York: CRC.

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

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

[24] Coolen, F.P.A. (2006). On probabilistic safety assessment in case of zero failures.*Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability*, **220**, 105-114.

[25] Coolen, F.P.A. (2010). Nonparametric predictive inference. In: *International Encyclopedia of Statistical Science*, Miodrag Lovric (Ed.), Springer, Berlin, 968–970.

[26] Coolen, F.P.A. and Al-nefaiee, A.H. (2012). Nonparametric predictive inference for failure times of systems with exchangeable components. *Journal of Risk and Reliability.* **226**: 262–273.

[27] Coolen, F.P.A. and Augustin, T. (2005). Learning from multinomial data: a nonparametric predictive alternative to the Imprecise Dirichlet Model. *Proceedings of the Fourth International Symposium on Imprecise Probabilities and Their Applications*, F.G. Cozman, R. Nau and T. Seidenfeld (Eds), published by SIPTA, **5**,125-134.

[28] Coolen, F.P.A. and Coolen-Schrijner, P. (2005). Nonparametric predictive reliability demonstration for failure-free periods. *IMA Journal of Management Mathematics*, **16**, 1–11.

[29] Coolen, F.P.A. and Coolen-Schrijner, P. (2006). Nonparametric predictive subset selection for proportions. *Statistics and Probability Letters*, **76**, 1675-1684.

[30] Coolen-Schrijner, P., Coolen, F.P.A. and Shaw, S.C. (2006). Nonparametric adaptive opportunity-based age replacement strategies. *Journal of the Operational Research Society*, **57**, 63–81.

[31] 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**, pp.1-22.

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

[33] 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.

[34] Coolen, F.P.A. and Maturi, T.A. (2010). Nonparametric predictive inference for order statistics of future observations. In: *Combining Soft Computing and Statistical Methods in Data Analysis.* C. Borgelt, *et al.* (eds). Springer, Berlin (Advances in Intelligent and Soft Computing 77, 97-104.

[35] Coolen, F.P.A. and Van der Laan, P. (2001). Imprecise predictive selection based on low structure assumptions. *Journal of Statistical Planning and Inference.* **98**, 259-277.

[36] Coolen, F.P.A. and Yan, K.J. (2003). Nonparametric predictive comparison of two groups of lifetime data. In Proceedings of the Third International Symposium on Imprecise Probabilities and Their Applications, in Proceedings *ISIPTA*. **3**, 148-161.

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

[38] Coolen-Maturi, T., Coolen, F.P.A. and Muhammad, N. (2016). Predictive inference for bivariate data: Combining nonparametric predictive inference for marginals with an estimated copula. *Journal of Statistical Theory and Practice*, **10**, 515-538.

[39] Das, K. R. and Imon, A. H. M. R. (2016). A brief review of tests for normality. *American Journal of Theoretical and Applied Statistics*, **5**, 5-12.

[40] David, H.A. and Nagaraja, H.N. (2003). *Order Statistics.* New Jersey: Wiley & Sons.

[41] Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge University Press.

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

[43] De Capitani, L. and De Martini, D. (2011). On stochastic orderings of the Wilcoxon rank sum test statistic- With applications to reproducibility probability estimation testing. *Statistics and Probability Letter*, **81**, 937–946.

[44] De Capitani, L. and De Martini, D. (2015). Reproducibility probability estimation and RP-testing for some nonparametric tests. *Entropy*, **18**,142.

[45] De Finetti, B. (1974). Theory of Probability. London: Wiley.

[46] Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P. and Meester, L. E. (2005). *A Modern Introduction to Probability and Statistics: Understanding Why and How*. London: Springer Science and Business Media.

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

[48] Diana, G. and Perri, P.F. (2011). A class of estimators for quantitative sensitive data. Statistical Papers, **52**, 633-650.

[49] Dixon, W. J. and Mood, A. M. (1946). The statistical sign test. *Journal of the American Statistical Association*, **41**, 557-566.

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

[51] Eichhorn, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference*, **7**, 307-316.

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

[53] Eriksson, S. A. (1973). A new model for randomized response. *International Statistical Review*, **41** 101-113.

[54] Fleiss, J. L., Levin, B. and Paik, M. C. (2003). *Statistical Methods for Rates and Proportions*, Third Edition, New York: John Wiley and Sons.

[55] Folsom, R. E., Greenberg, B. G., Horvitz, D. G. and Abernathy, J. R. (1973). The two alternate questions randomized response model for human surveys. *Journal of the American Statistical Association*, **68**, 525-530.

[56] Geisser, S. (1993). *Predictive Inference: An Introduction.* London: Chapman and Hall.

[57] Gelman, A. (2015). The connection between varying treatment effects and the crisis of unreliable research: A Bayesian perspective, **41**, 632-643.

[58] Gibbons, J.D. and Chakraborti, S. (2011). *Nonparametric Statistical Inference (5th ed.).* Chapman and Hall, Boca Raton, Florida.

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

[60] Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, **64**, 520-539.

[61] Greenberg, B. G., Kuebler, R. R., Abernathy, J. R. and Horvitz, D. G. (1971). Application of the randomized response technique in obtaining quantitative data,*Journal of American Statistical Association*, **66**, 243-250.

[62] Gupta, S.N., Gupta, B.C. and Singh, S.(2002). Estimation of sensitivity level of personal interview survey questions, *Journal of Statistical Planning and Inference*, **100**, 239-247.

[63] Gupta, S., Mehta, S., Shabbir, J. and Khalil, S. (2018). A unified measure of respondent privacy and model efficiency in quantitative RRT models. *Journal of Statistical Theory and Practice*, **12**, 506-511.

[64] Gupta, S.N. and Shabbir, J. (2004), Sensitivity estimation for personal interview survey questions, *Statistica*, **64**, 643-653.

[65] Gupta, S., Shabbir, J. and Sehra, S. (2010). Mean and sensitivity estimation in optional randomized response models. *Journal of Statistical Planning and Inference*, **140**, 2870-2874.

[66] Gupta, S. and Thornton, B. (2002). Circumventing social desirability response bias in personal interview surveys. *American Journal of Mathematical and Management Sciences*, **22**, 369-383.

[67] Gupta, S. N., Thornton, B., Shabbir, J. and Singhal, S. (2006). A comparison of multiplicative and additive optional RRT models, *Journal of Statistical Theory and Applications*, **5**, 226-239.

[68] He, T., Coolen, F.P.A. and Coolen-Maturi, T. (2019). Nonparametric predictive inference for European option pricing based on the Binomial Tree Model. *Journal of the Operational Research Society.* **70**, 1692-1708.

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

[70] Hill, B.M. (1988). De Finetti's Theorem, Induction, and $A_{(n)}$ or Bayesian nonparametric predictive inference (with discussion). In J.M. Bernardo, et al. (Eds.), *Bayesian Statistics*, **3**, 211-241. Oxford University Press.

[71] Hussain, Z., Al-Sobhi, M. M. and Al-Zahrani, B. (2014). Additive and subtractive scrambling in optional randomized response modeling. *PLoS One*, **6**, e83557.

[72] Jones, O., Maillardet, R. and Robinson, A. (2014). *Introduction to Scientific Programming and Simulation Using R.* New York: Chapman and Hall/CRC.

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

[74] Kokoska, S. and Nevison, C. (1989) Cumulative Distribution Function for the Standard Normal Random Variable. In: *Statistical Tables and Formulae.* 55–56.

[75] Kuk, A. Y. (1990). Asking sensitive questions indirectly. *Biometrika*, **77**, 436-438.

[76] Lanke, J. (1976). On the degree of protection in randomized interviews. *International Statistical Review*, **44**, 197-203.

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

[78] Linacre, J. (1996) Overlapping Normal Distributions. *Transactions of the Rasch Measurement.* **10**, 487.

[79] Ljungqvist, L. (1993). A unified approach to measures of privacy in randomized response models: A utilitarian perspective. *Journal of the American Statistical Association*, **88**, 97-103

[80] Mangat, N. S. and Singh, R. (1990). An alternative randomized response procedure. Biometrika, **77**, 439-442.

[81] Marques, F.J. Coolen, F.P.A. and Coolen-Maturi, T. (2019). Approximations for the likelihood ratio statistic for hypothesis testing between two Beta distributions. *Journal of Statistical Theory and Practice*, **13**, 17.

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

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

[84] Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model. *Journal of the American Statistical Association.* **66**, 627-629.

[85] Odumade, O. and Singh, S.(2009). Efficient use of two decks of cards in randomized response sampling. *Communications in Statistics-Theory and Methods*, **38**, 439-446.

[86] Pastore, M. and Calcagnì, A. (2019). Measuring distribution similarities between samples: A distribution-free overlapping index. Frontiers in Psychology, **10**, 1089.

[87] Posavac, E. J. (2002). Using p values to estimate the probability of a statistically significant replication. *Understanding Statistics: Statistical Issues in Psychology*, Education, and the Social Sciences, **1**, 101–112.

[88] Reiser, B. and Faraggi, D. (1999). Confidence intervals for the overlapping coefficient: the normal equal variance case. *Journal of the Royal Statistical Society: Series D (the Statistician).* **48**, 413-418.

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

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

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

[92] Thulin, M. (2014). The cost of using exact confidence intervals for a binomial proportion. *Electronic Journal of Statistics*, **8**, 817-840.

[93] Vollset, S. E. (1993). Confidence intervals for a binomial proportion. *Statistics in Medicine*, **12**, 809-824.

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

[95] Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63–69.

[96] Weichselberger, K.(2000). The theory of interval-probability as a unifying concept for uncertainty. *International Journal of Approximate Reasoning*, **24**, 149-170.

[97] Weichselberger, K. (2001). Elementare Grundbegriffe einer Allgemeineren Wahrscheinlichkeitsrechnung I. Intervallwahrscheinlichkeit als Umfassendes Konzept (In German). Physika, Heidelberg.

[98] Yan, Z., Wang, J. and Lai, J. (2008). An efficiency and protection degree-based comparison among the quantitative randomized response strategies. *Communications in Statistics-Theory and Methods*, **38**, 400-408.

[99] Young, A., Gupta, S. and Parks, R. (2019). A binary unrelated-question RRT model accounting for untruthful responding. *Involve: A Journal of Mathematics*, **12**(7), 1163-1173.

[100] Zhimin, H. and Zaizai, Y. (2012). Measure of privacy in randomized response model. *Quality and Quantity*, **46**, 1167-1180.